(D4)

⑲ Europäisches Patentamt

European Patent Office

Office européen des brevets

⑪ Publication number: **0 590 747 A2**

⑫ **EUROPEAN PATENT APPLICATION**

㉑ Application number: 93250268.5

㉒ Date of filing: 29.09.93

� Int. Cl.⁵: **C12N 15/54**, C12N 9/10,
C12N 15/12, C07K 13/00,
C12N 15/86, C12N 15/64,
C12N 5/10

The applicant has subsequently filed a sequence listing and declared, that it includes no new matter.

㉚ Priority: **01.10.92 US 955041**

㊸ Date of publication of application:
06.04.94 Bulletin 94/14

㊹ Designated Contracting States:
AT BE CH DE DK ES FR GB GR IE IT LI LU MC NL PT SE

㉛ Applicant: **LA JOLLA CANCER RESEARCH FOUNDATION**
10901 North Torrey Pines Road
La Jolla, CA 92037(US)

㉜ Inventor: **Fukuda, Minoru**
2818 Passy Avenue
San Diego, California 92122(US)
Inventor: **Bierhuizen, Marti F. A.**
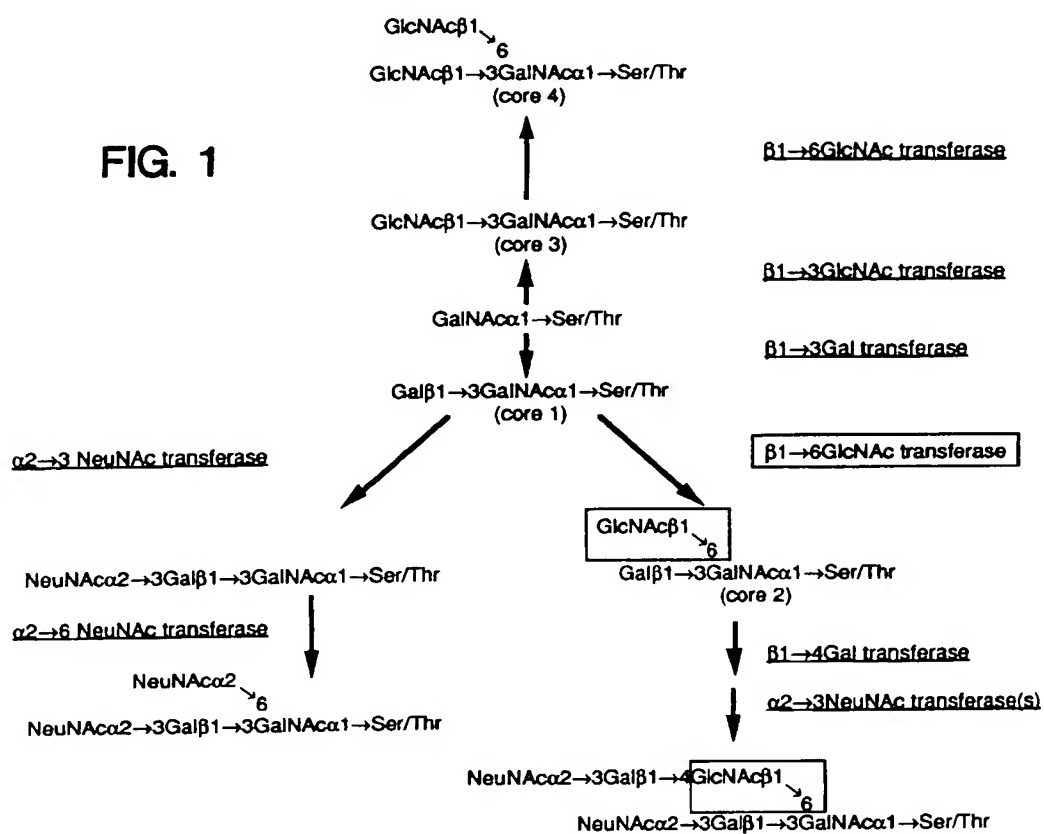**Schiedamseweg 225**
NL-3119 JB Schiedam(US)

㉞ Representative: **UEXKÜLL & STOLBERG**
Patentanwälte
Beselerstrasse 4
D-22607 Hamburg (DE)

㊼ A novel-beta1- 6 N-acetylglucosaminyltransferase, its acceptor molecule, leukosialin, and a method for cloning proteins having enzymatic activity.

�57 The present invention provides a novel $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase, which forms core 2 oligosaccharide structures in O-glycans, and a novel acceptor molecule, leukosialin, CD43, for core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase activity. The amino acid sequences and nucleic acid sequences encoding these molecules, as well as active fragments thereof, also are disclosed. A method for isolating nucleic acid sequences encoding proteins having enzymatic activity is disclosed, using CHO cells that support replication of plasmid vectors having a polyoma virus origin of replication. A method to obtain a suitable cell line that expresses an acceptor molecule also is disclosed.

EP 0 590 747 A2

GlcNAcβ1
       ↘6
GlcNAcβ1→3GalNAcα1→Ser/Thr
(core 4)

FIG. 1

β1→6GlcNAc transferase

GlcNAcβ1→3GalNAcα1→Ser/Thr
(core 3)

β1→3GlcNAc transferase

GalNAcα1→Ser/Thr

β1→3Gal transferase

Galβ1→3GalNAcα1→Ser/Thr
(core 1)

α2→3 NeuNAc transferase

β1→6GlcNAc transferase

GlcNAcβ1
      ↘6

NeuNAcα2→3Galβ1→3GalNAcα1→Ser/Thr

Galβ1→3GalNAcα1→Ser/Thr
(core 2)

α2→6 NeuNAc transferase

β1→4Gal transferase

NeuNAcα2
    ↘6
NeuNAcα2→3Galβ1→3GalNAcα1→Ser/Thr

α2→3NeuNAc transferase(s)

NeuNAcα2→3Galβ1→4GlcNAcβ1
                     ↘6
NeuNAcα2→3Galβ1→3GalNAcα1→Ser/Thr

BACKGROUND OF THE INVENTION

5

FIELD OF THE INVENTION

This invention relates generally to the fields of biochemistry and molecular biology and more specifically to a novel human enzyme, UDP-GlcNAc:Galβ1→3GalNAc (GlcNAc to GalNAc) β1→6 N-acetyl

10 glucosaminyltransferase (core 2 β1→6 N-acetylglucosaminyltransferase; C2GnT), and to a novel acceptor molecule, leukosialin, CD43, for core 2 β1→6 N-acetylglucosaminyltransferase action. The invention additionally relates to DNA sequences encoding core 2 β1→6 N-acetylglucosaminyltransferase and leukosialin, to vectors containing a C2GnT DNA sequence or a leukosialin DNA sequence, to recombinant host cells transformed with such vectors and to a method of transient expression cloning in CHO cells for identifying

15 and isolating DNA sequences encoding specific proteins, using CHO cells expressing a suitable acceptor molecule.

BACKGROUND INFORMATION

20 Most O-glycosidic oligosaccharides in mammalian glycoproteins are linked via N-acetylgalactosamine to the hydroxyl groups of serine or threonine. These O-glycans can be classified into 4 different groups depending on the nature of the core portion of the oligosaccharides (see Fig. 1). Although less well studied than N-glycans, O-glycans likely have important biological functions. Indeed, the presence of O-linked oligosaccharides with the core 2 branch, Galβ1→3(GlcNAcβ1→6)GalNAc, has been demonstrated in many

25 biological processes.

Piller et al., J. Biol. Chem 263:15146-15150 (1988) reported that human T-cell activation is associated with the conversion of core 1-based tetrasaccharides to core 2-based hexasaccharides on leukosialin, a major sialoglycoprotein present on human T lymphocytes (see also Fig. 1). A similar increase in hexasaccharides was observed in peripheral blood lymphocytes of patients suffering from T-cell leukemias (Saitoh

30 et al., Blood 77:1491-1499 (1991)), myelogenous leukemias (Brockhausen et al., Cancer Res. 51:1257-1263 (1991)) and immunodeficiency due to AIDS and the Wiskott-Aldrich syndrome (Piller et al., J. Exp. Med. 173:1501-1510 (1991)). In these patients' lymphocytes, changes in the amount of hexasaccharides were caused by increased activity of either UDP-GlcNAc:Galβ1→3GalNAc (GlcNAc to GalNAc) 6-β-D-N-acetylglucosaminyltransferase (EC2.4.1.102) or core 2 β1→6 N-acetylglucosaminyltransferase (Williams et al., J.

35 Biol. Chem. 255:11253-11261 (1980)). Increased activity of core 2 β1→6 N-acetylglucosaminyltransferase also was observed in metastatic murine tumor cell lines as compared to their parental, non-metastatic counterparts (Yousefi et al., J. Biol. Chem. 266:1772-1782 (1991)).

Increased complexity of the attached oligosaccharides increases the molecular weight of the glycoprotein. For example, leukosialin containing hexasaccharides has a molecular weight of ~135kDa,

40 whereas leukosialin containing tetrasaccharides has a molecular weight of ~105kDa (Carlsson et al., J. Biol. Chem. 261:12779-12786 and 12787-12795 (1986)).

Fox et al., J. Immunol. 131:762-767 (1983) raised a monoclonal antibody, T305, against human T-lymphocytic leukemia cells. Sportsman et al., J. Immunol. 135:158-164 (1985) reported T305 binding was abolished by neuraminidase treatment, suggesting T305 binds to hexasaccharides. T305 specifically reacts

45 with the high molecular weight form of leukosialin (Saitoh et al., supra, (1991)).

Previous studies indicated poly-N-acetyllactosamine repeats extend almost exclusively from the branch formed by the core 2 β1→6 N-acetylglucosaminyltransferase (Fukuda et al., J. Biol. Chem. 261:12796-12806 (1986)). Consistent with these results, Yousefi et al., supra, (1991) demonstrated that the core 2 enzyme in metastatic tumor cells regulates the level of poly-N-acetyllactosamine synthesis in O-linked

50 oligosaccharides.

Poly-N-acetyllactosamines are subject to a variety of modifications, including the formation of the sialyl Le^X, NeuNAcα2→3Galβ1→4(Fucα1→3)GlcNAc-, or the sialyl Le^a, NeuNAcα2→3Galβ1→3 (Fucα1→4)GlcNAc, determinants (Fukuda, Biochim. Bionhys. Acta 780:119-150 (1985)). Such modifications are significant because these determinants, which are present on neutrophils and monocytes, serve as ligands for E- and

55 P-selectin present on endothelial cells and platelets, respectively (see, for example, Larsen et al., Cell 63:467-474 (1990)).

In addition, tumor cells often express a significant amount of sialyl Le^X and/or sialyl Le^a on their cell surfaces. The interaction between E-selectin or P-selectin and these cell surface carbohydrates may play a

role in tumor cell adhesion to endothelium during the metastatic process (Walz et al., supra, (1990)). Kojima et al., Biochem. Biophys. Res. Commun. 182:1288-1295 (1992) reported that selectin-dependent tumor cell adhesion to endothelial cells was abolished by blocking O-glycan synthesis. Complex sulfated O-glycans also may serve as ligands for the lymphocyte homing receptor, L-selectin (Imai et al., J. Cell Biol. 113:1213-
5   1221 (1991)).

These reported observations establish core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase as a critical enzyme in O-glycan biosynthesis. The availability of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase will allow the in vivo and in vitro production of specific glycoproteins having core 2 oligosaccharides and subsequent study of these variant O-glycans on cell-cell interactions. For example, core 2 $\beta1\rightarrow6$ N-
10 acetylglucosaminyltransferase is a useful marker for transformed or cancerous cells. An understanding of the role of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase in transformed and cancerous cells may elucidate a mechanism for the aberrant cell-cell interactions observed in these cells. In order to understand the control of expression of these oligosaccharides and their function, isolation of a cDNA clone for core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase is a prerequisite. However, the DNA sequence encoding core 2
15 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase has not yet been reported.

Thus, a need exists for identifying the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase and the DNA sequences encoding this enzyme. The present invention satisfies this need and provides related advantages as well.

20 **SUMMARY OF THE INVENTION**

The present invention generally relates to a novel purified human $\beta1\rightarrow6$ N-acetylglucosaminyltransferase. A cDNA sequence encoding a 428 amino acid protein having $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity also is provided. The purified human $\beta1\rightarrow6$ N-acetylglucosaminyltransferase, or an active fragment
25 thereof, catalyzes the formation of critical branches in O-glycans.

The invention further relates to a novel purified acceptor molecule, leukosialin, CD43, for core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity. The leukosialin cDNA encodes a novel variant leukosialin, which is created by alternative splicing of the genomic leukosialin DNA sequence.

Isolated nucleic acids encoding either core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase or leukosialin are
30 disclosed, as are vectors containing the nucleic acids and recombinant host cells transformed with such vectors. The invention further provides methods of detecting such nucleic acids by contacting a sample with a nucleic acid probe having a nucleotide sequence capable of hybridizing with the isolated nucleic acids of the present invention. The core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase and leukosialin amino acid and nucleic acid sequences disclosed herein can be purified from human cells or produced using well known
35 methods of recombinant DNA technology.

The invention also discloses a method of isolating nucleic acid sequences encoding proteins that have an enzymatic activity. Such a nucleic acid sequence is obtained by transfecting the nucleic acid, which is contained within a vector having a polyoma virus replication origin, into a Chinese hamster ovary (CHO) cell line simultaneously expressing polyoma virus large T antigen and the acceptor molecule for the protein
40 having an enzymatic activity.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 depicts the structures and biosynthesis of O-glycans. Structures of O-glycan cores can be
45 classified into 4 groups (core 1 to core 4), each of which is synthesized starting with GalNAc$\alpha1\rightarrow$Ser/Thr. The core 1 structure is synthesized by the addition of a $\beta1\rightarrow3$ Gal residue to the GalNAc residue. The core 1 structure can be converted to core 2 by the addition of a $\beta1\rightarrow6$ N-acetylglucosaminyl residue. This intermediate is usually converted to the hexasaccharide by sequential addition of galactose and sialic acid residues (bottom right). The core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase and the linkage formed by the
50 enzyme are indicated by a box. In certain cell types, the core 2 structure can be extended by the addition of N-acetyllactosamine (Gal$\beta1\rightarrow4$GlcNAc$\beta1\rightarrow3$) repeats to form poly-N-acetyllactosamine. In the absence of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase, core 1 is converted to the monosialoform, then to the disialoform by sequential addition of $\alpha2\rightarrow3$- and $\alpha2\rightarrow6$-linked sialic acid residues (bottom left). Alternatively, core 3 can be synthesized by the addition of a $\beta1\rightarrow3$ N-acetylglucosaminyl residue to the GalNAc residue.
55 Core 3 can be converted to core 4 by another $\beta1\rightarrow6$ N-acetylglucosaminyltransferase (top of figure).

Figure 2 depicts genomic DNA sequence (SEQ. ID. NO. 1) and cDNA sequence (SEQ. ID. NO. 1) of leukosialin. The genomic sequence is numbered relative to the transcriptional start site. Exon 1 and exon 2 have been previously described. Exon 1' is newly identified here. In the isolated cDNA, exon 1' is

immediately followed by the exon 2 sequence. Deduced amino acids (SEQ. ID. NO. 2) are presented under the coding sequence, which begins in exon 2. A portion of the exon 2 sequence is shown.

Figure 3 establishes the ability of pGT/hCG to replicate in CHO cell lines expressing polyoma large T antigen and leukosialin. In panel A, six clonal CHO cell lines were examined for replication of pcDNAI-based pGT/hCG (lanes 1-6). In panel B, replication of cell clone 5 (CHO-Py-leu), was further examined by treatment with increasing concentrations of DpnI and XhoI (lanes 2 and 3). Plasmid DNA isolated from MOP-8 cells was used as a control (lane 1). Plasmid DNA was extracted using the Hirt procedure and samples were digested with XhoI and DpnI. In parallel, pGT/hCG plasmid purified from *E. coli* MC1061/P3 was digested with XhoI and DpnI (lane 7 in panel A and lane 4 in panel B) or XhoI alone (lane 8 in panel A and lane 5 in panel B). The arrow indicates the migration of plasmid DNA resistant to DpnI digestion. The arrowheads indicate plasmid DNA digested by DpnI.

Figure 4 shows the expression of T305 antigen expressed by pcDNAI-C2GnT. Subconfluent CHO-Py-leu cells were transfected with pcDNAI-C2GnT (panels A and B) or mock-transfected with pcDNAI (panels C and D). Sixty four hours after transfection, the cells were fixed, then incubated with mouse T305 monoclonal antibody followed by fluorescein isocyanate-conjugated sheep anti-mouse IgG (panels A, B and C). Two different areas are shown in panels A and B. Panel D shows a phase micrograph of the same field shown in panel C. Bar = 20μm.

Figure 5 depicts the cDNA sequence (SEQ. ID. NO. 3) and translated amino acid sequences (SEQ. ID. NO. 4) of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase The open reading frame and full-length nucleotide sequence of C2GnT are shown. The signal/membrane-anchoring domain is doubly underlined. The polyadenylation signal is boxed. Potential N-glycosylation sites are marked with asterisks. The sequences are numbered relative to the translation start site.

Figure 6 shows the expression of core 2 $\beta1\rightarrow6$ N−acetylglucosaminyltransferase mRNA in various cell types. Poly(A)$^+$ RNA (11 μg) from CHO-Py-leu cells (lane 1), HL-60 promyelocytes (lane 2), K562 erythrocytic cells (lane 3), and SP and L4 colonic carcinoma cells (lanes 4 and 5) was resolved by electrophoresis. RNA was transferred to a nylon membrane and hybridized with a radiolabeled fragment of pPROTA-C2GnT. Migration of RNA size markers is indicated.

Figure 7 illustrates the construction of the vector encoding the protein A-C2GnT fusion protein. The cDNA sequence corresponding to Pro^{38} to His^{428} was fused in frame with the IgG binding domain of *S. aureus* protein A (bottom; SEQ. ID. NOS. 7 and 8). The sequence includes the cleavable signal peptide, which allows secretion of the fused protein. The coding sequence is under control of the SV40 promoter. The remainder of the vector sequence shown was derived from rabbit β-globin gene sequences, including an intervening sequence (IVS) and a polyadenylation signal (An).

DETAILED DESCRIPTION OF THE INVENTION

The present invention generally relates to a novel human core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase. The invention further relates to a novel method of transient expression cloning in CHO cells that was used to isolate the cDNA sequence encoding human core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase (C2GnT). The invention also relates to a novel human leukosialin, which is an acceptor molecule for core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity.

Cells generally contain extremely low amounts of glycosyltransferases. As a result, cDNA cloning based on screening using an antibody or a probe based on the glycosyltransferase amino acid sequence has met with limited success. However, isolation of cDNAs encoding various glycosyltransferases can be achieved by transient expression of cDNA in recipient cells.

Successful application of the transient expression cloning method to isolate a cDNA sequence encoding a glycosyltransferase requires an appropriate recipient cell line. Ideal recipient cells should not express the glycosyltransferase of interest. As a result, the recipient cells would normally lack the oligosaccharide structure formed by such a glycosyltransferase.

Expression of the cloned glycosyltransferase cDNA in the recipient cell line should result in formation of the specific oligosaccharide structure. The resultant oligosaccharide can be identified using a specific antibody or lectin that recognizes the structure. The recipient cell line also must support replication of an appropriate plasmid vector.

COS-1 cells initially appear to satisfy the requirements for using the transient expression method. COS-1 cells express SV40 large T antigen and support the replication of plasmid vectors harboring a SV40 replication origin (Gluzman et al., Cell 23:175-182 (1981)). Although COS-1 cells, themselves, express a variety of glycosyltransferases, COS-1 cells have been used to clone cDNA sequences encoding human blood group Lewis $\alpha1\rightarrow3/4$ fucosyltransferase and murine $\alpha1\rightarrow3$ galactosyltransferase (Kukowska-Latallo et

al., Genes and Devel. 4:1288-1303 (1990); Larsen et al., Proc. Natl. Acad. Sci. USA 86:8227-8231 (1989)). Also, Goelz et al., Cell 63:175-182 (1990), utilized an antibody that inhibits E-selectin mediated adhesion to isolate a cDNA sequence encoding $\alpha 1 \rightarrow 3$ fucosyltransferase.

5   An attempt was made to use COS-1 cells to isolate cDNA clones encoding core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase. COS-1 cells were transfected using cDNA obtained from activated human T cells, which express the core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase. Transfected cells suspected of expressing core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase in the transfected cells were identified by the presence of increased levels of the core 2 oligosaccharide structure formed by core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase activity. The presence of the core 2 structure was identified using the 10   monoclonal antibody, T305, which identifies a hexasaccharide on leukosialin. A clone expressing high levels of the T305 antigen was isolated and sequenced.

    Surprisingly, transfection using COS-1 cells resulted in the isolation of a cDNA clone encoding a novel variant of human leukosialin, which is the acceptor molecule for core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase activity. Examination of the cDNA sequence of the newly isolated leukosialin revealed the cDNA 15   sequence was formed as a result of alternative splicing of exons in the genomic leukosialin DNA sequence. Specifically, the newly isolated leukosialin is encoded by cDNA sequence containing a previously un-described non-coding exon at the 5'-terminus (exon 1' in Figure 2; SEQ. ID. NO. 1).

    The unexpected result obtained using COS-1 cells led to the development of a new transfection system to isolate a cDNA sequence encoding core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase. CHO cells, which do 20   not normally express the T305 antigen, were transfected with DNA sequences encoding human leukosialin and the polyoma virus large T antigen. A cell line, designated CHO-Py-leu, which expresses human leukosialin and polyoma virus large T antigen, was isolated.

    CHO-Py-leu cells were used for transient expression cloning of a cDNA sequence encoding core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase. CHO-Py-leu cells were transfected with cDNA obtained from human 25   HL-60 promyelocytes A plasmid, pcDNAI-C2Gnt, which directed expression of the T305 antigen, was isolated and the cDNA insert was sequenced (see Figure 5; SEQ. ID. NO. 3). The 2105 base pair cDNA sequence encodes a putative 428 amino acid protein (SEQ. ID. NO. 4). The genomic DNA sequence encoding can be isolated using methods well known to those skilled in the art, such as nucleic acid hybridization using the core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase cDNA disclosed herein to screen, for 30   example, a genomic library prepared from HL-60 promyelocytes.

    An enzyme similar to the disclosed human core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase has been purified from bovine tracheal epithelium (Ropp et al., J. Biol. Chem. 266:23863-23871 (1991), which is incorporated herein by reference. The apparent molecular weight of the bovine enzyme is ~69kDa. In comparison, the predicted molecular weight of the polypeptide portion of core 2 $\beta 1 \rightarrow 6$ N-acetyl-35   glucosaminyltransferase is ~50kDa. The deduced amino acid sequence of core 2 $\beta 1 \rightarrow 6$ N-acetyl-glucosaminyltransferase reveals two to three potential N-glycosylation sites, suggesting N-glycosylation and O-glycosylation, or other post-translational modification, could account for the larger apparent size of the bovine enzyme.

    Expression of the cloned C2GnT sequence, or a fragment thereof, directed formation of the specific O-40   glycan core 2 oligosaccharide structure. Although several cDNA sequences encoding glycosyltransferases have been isolated (Paulson and Colley, J. Biol. Chem. 264:17615-17618 (1989); Schachter, Curr. Opin. Struct. Biol. 1:755-765 (1991), which are incorporated herein by reference), C2GnT is the first reported cDNA sequence encoding an enzyme involved exclusively in O-glycan synthesis.

    In O-glycans, $\beta 1 \rightarrow 6$ N-acetylglucosaminyl linkages may occur in both core 2, Gal$\beta 1 \rightarrow 3$(GlcNAc$\beta 1 \rightarrow 6$)-45   GalNAc, and core 4, GlcNAc$\beta 1 \rightarrow 3$(GlcNAc$\beta 1 \rightarrow 6$)GalNAc, structures (Brockhausen et al., Biochemistry 24:1866-1874 (1985), which is incorporated herein by reference. In addition, $\beta 1 \rightarrow 6$ N-acetylglucosaminyl linkages occur in the side chains of poly-N-acetyllactosamine, forming the I-structure (Piller et al., J. Biol. Chem. 259:13385-13390 (1984), which is incorporated herein by reference), and in the side chain attached to α-mannose of the N-glycan core structure, forming a tetraantennary saccharide (Cummings et al., J. Biol. 50   Chem. 257:13421-13427 (1982), which is incorporated herein by reference). The enzymes responsible for these linkages all share the unique property that Mn^{2+} is not required for their activity.

    Although it was originally suggested that these $\beta 1 \rightarrow 6$ N-acetylglucosaminyl linkages were formed by the same enzyme (Piller at al., 1984), the present disclosure clearly demonstrates that the HL-60-derived core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase is specific for the formation only of O-glycan core 2. This 55   result is consistent with a recent report demonstrating that myeloid cell lysates contain the enzymatic activity associated with core 2, but not core 4, formation (Brockhausen et al., supra, (1991)).

    Analysis of mRNA isolated from colonic cancer cells indicated core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltrans-ferase is expressed in these cells. Recent studies using affinity absorption suggested at least two different

$\beta1\rightarrow6$ N-acetylglucosaminyltransferases were present in tracheal epithelium (Ropp et al., supra, (1991)). One of these transferases formed core 2, core 4, and I structures. Thus, at least one other $\beta1\rightarrow6$ N-acetylglucosaminyltransferase present in epithelial cells can form core 2, core 4 and I structures. Similarly, a $\beta1\rightarrow6$ N-acetylglucosaminyltransferase present in Novikoff hepatoma cells can form both core 2 and I
5   structures (Koenderman et al., Eur. J. Biochem. 166:199-208 (1987), which is incorporated herein by reference).

The acceptor molecule specificity of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase is different from the specificity of the enzymes present in tracheal epithelium and Novikoff hepatoma cells. Thus, a family of $\beta1\rightarrow6$ N-acetylglucosaminyltransferases can exist, the members of which differ in acceptor specificity but
10   are capable of forming the same linkage. Members of this family are isolated from cells expressing $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity, using, for example, nucleic acid hybridization assays and studies of acceptor molecule specificity. Such a family was reported for the $\alpha1\rightarrow3$ fucosyltransferases (Weston et al., J. Biol. Chem. 267:4152-4160 (1992), which is incorporated herein by reference).

The formation of the core 2 structure is critical to cell structure and function. For example, the core 2
15   structure is essential for elongation of poly-N-acetyllactosamine and for formation of sialyl Le^X or sialyl Le^a structures. Furthermore, the biosynthesis of cartilage keratan sulfate may be initiated by the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase, since the keratan sulfate chain is extended from a branch present in core 2 structure in the same way as poly-N-acetyllactosamine (Dickenson et al., Biochem. J. 269:55-59 (1990), which is incorporated herein by reference). Keratan sulfate is absent in wild-type CHO cells, which do not
20   express the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase (Esko et al., J. Biol. Chem. 261:15725-15733 (1986), which is incorporated herein by reference). These structures are believed to be important for cellular recognition and matrix formation. The availability of the cDNA clone encoding the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase will aid in understanding how the various carbohydrate structures are formed during differentiation and malignancy. Manipulation of the expression of the various carbohydrate structures
25   by gene transfer and gene inactivation methods will help elucidate the various functions of these structures.

The present invention is directed to a method for transient expression cloning in CHO cells of cDNA sequences encoding proteins having enzymatic activity. Isolation of human core 2 $\beta1\rightarrow6$ N-acetyl-glucosaminyltransferase is provided as an example of the disclosed method. However, the method can be used to obtain cDNA sequences encoding other proteins having enzymatic activity.
30   For example, lectins and antibodies reactive with other specific oligosaccharide structures are available and can be used to screen for glycosyltransferase activity. Also, CHO cell lines that have defects in glycosylation have been isolated. These cell lines can be used to study the activity of the corresponding glycosyltransferase (Stanley, Ann. Rev. Genet. 18:525-552 (1984), which is incorporated herein by reference). CHO cell lines also have been selected for various defects in cellular metabolism, loss of expression
35   of cell surface molecules and resistance to cytotoxic drugs (see, for example, Malmström and Krieger, J. Biol. Chem. 266:24025-24030 (1991); Yayon et al., Cell 64:841-848 (1991), which are incorporated herein by reference). The approach disclosed herein should allow isolation of cDNA sequences encoding the proteins involved in these various cellular functions.

As used herein, the terms "purified" and "isolated" mean that the molecule or compound is substan-
40   tially free of contaminants normally associated with a native or natural environment. For example, a purified protein can be obtained from a number of methods. The naturally-occurring protein can be purified by any means known in the art, including, for example, by affinity purification with antibodies having specific reactivity with the protein. In this regard, anti-core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase antibodies can be used to substantially purify naturally-occurring core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase from
45   human HL-60 promyelocytes.

Alternatively, a purified protein of the present invention can be obtained by well known recombinant methods, utilizing the nucleic acids disclosed herein, as described, for example, in Sambrook et al., Molecular Cloning: A Laboratory Manual 2d ed. (Cold Spring Harbor Laboratory 1989), which is incorporated herein by reference, and by the methods described in the Examples below. Furthermore, purified proteins
50   can be synthesized by methods well known in the art.

As used herein, the phrase "substantially the sequence" includes the described nucleotide or amino acid sequence and sequences having one or more additions, deletions or substitutions that do not substantially affect the ability of the sequence to encode a protein have a desired functional activity. In addition, the phrase encompasses any additional sequence that hybridizes to the disclosed sequence under
55   stringent hybridization sequences. Methods of hybridization are well known to those skilled in the art. For example, sequence modifications that do not substantially alter such activity are intended. Thus, a protein having substantially the amino acid sequence of Figure 5 (SEQ. ID. NO. 4) refers to core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase encoded by the cDNA described in Example IV, as well as proteins having

amino acid sequences that are modified but, nevertheless, retain the functions of core 2 β1→6 N-acetylglucosaminyltransferase. One skilled in the art can readily determine such retention of function following the guidance set forth, for example, in Examples V and VI.

5   The present invention is further directed to active fragments of the human core 2 β1→6 N-acetylglucosaminyltransferase protein. As used herein, an active fragment refers to portions of the protein that substantially retain the glycosyltransferase activity of the intact core 2 β1→6 N-acetylglucosaminyltransferase protein. One skilled in the art can readily identify active fragments of proteins such as core 2 β1→6 N-acetylglucosaminyltransferase by comparing the activities of a selected fragment with the intact protein following the guidance set forth in the Examples below.

10   As used herein, the term "glycosyltransferase activity" refers to the function of a glycosyltransferase to link sugar residues together through a glycosidic bond to create critical branches in oligosaccharides. Glycosyltransferase activity results in the specific transfer of a monosaccharide to an appropriate acceptor molecule, such that the acceptor molecule contains oligosaccharides having critical branches. One skilled in the art would understand the terms "enzymatic activity" and "catalytic activity" to generally refer to a

15   function of certain proteins, such as the function of those proteins having glycosyltransferase activity.

As used herein, the term "acceptor molecule" refers to a molecule that is acted upon by a protein having enzymatic activity. For example, an acceptor molecule, such as leukosialin, as identified by the amino acid sequence of Figure 2 (SEQ. ID. NO. 2), accepts the transfer of a monosaccharide due to glycosyltransferase activity. An acceptor molecule, such as leukosialin, may already contain one or more

20   sugar residues. The transfer of monosaccharides to an acceptor molecule, such as leukosialin, results in the formation of critical branches of oligosaccharides.

As used herein, the term "critical branches" refers to oligosaccharide structures formed by specific glycosyltransferase activity. Critical branches may be involved in various cellular functions, such as cell-cell recognition. The oligosaccharide structure of a critical branch can be determined using methods well known

25   in the art, such as the method for determining the core 2 oligosaccharide structure, as described in Examples V and VI.

Relatedly, the invention also provides nucleic acids encoding the human core 2 β1→6 N-acetylglucosaminyltransferase protein and leukosialin protein described above. The nucleic acids can be in the form of DNA, RNA or cDNA, such as the novel C2GnT cDNA of 2105 base pairs identified in Figure 5

30   (SEQ. ID. NO. 3) or the novel leukosialin cDNA identified in Figure 2 (SEQ. ID. NO. 1), for example. Such nucleic acids can also be chemically synthesized by methods known in the art, including, for example, the use of an automated nucleic acid synthesizer.

The nucleic acid can have substantially the nucleotide sequence of C2GnT, identified in Figure 5 (SEQ. ID. NO. 3), or leukosialin identified in Figure 2 (SEQ. ID. NO. 1). Portions of such nucleic acids that encode

35   active fragments of the core 2 β1→6 N-acetylglucosaminyltransferase protein or leukosialin protein of the present invention also are contemplated.

Nucleic acid probes capable of hybridizing to the nucleic acids of the present invention under reasonably stringent conditions can be prepared from the cloned sequences or by synthesizing oligonucleotides by methods known in the art. The probes can be labeled with markers according to

40   methods known in the art and used to detect the nucleic acids of the present invention. Methods for detecting such nucleic acids can be accomplished by contacting the probe with a sample containing or suspected of containing the nucleic acid under hybridizing conditions, and detecting the hybridization of the probe to the nucleic acid.

The present invention is further directed to vectors containing the nucleic acids described above. The

45   term "vector" includes vectors that are capable of expressing nucleic acid sequences operably linked to regulatory sequences capable of effecting their expression. Numerous cloning vectors are known in the art. Thus, the selection of an appropriate cloning vector is a matter of choice. In general, useful vectors for recombinant DNA are often plasmids, which refer to circular double stranded DNA loops such as pcDNAI or pcDSRα. As used herein, "plasmid" and "vector" may be used interchangeably as the plasmid is a

50   common form of a vector. However, the invention is intended to include other forms of expression vectors that serve equivalent functions.

Suitable host cells containing the vectors of the present invention are also provided. Host cells can be transformed with a vector and used to express the desired recombinant or fusion protein. Methods of recombinant expression in a variety of host cells, such as mammalian, yeast, insect or bacterial cells are

55   widely known. For example, a nucleic acid encoding core 2 β1→6 N-acetylglucosaminyltransferase or a nucleic acid encoding leukosialin can be transfected into cells using the calcium phosphate technique or other transfection methods, such as those described in Sambrook et al., supra, (1989).

Alternatively, nucleic acids can be introduced into cells by infection with a retrovirus carrying the gene or genes of interest. For example, the gene can be cloned into a plasmid containing retroviral long terminal repeat sequences, the C2Gnt DNA sequence or the leukosialin DNA sequence, and an antibiotic resistance gene for selection. The construct can then be transfected into a suitable cell line, such as PA12, which

5 carries a packaging deficient provirus and expresses the necessary components for virus production, including synthesis of amphotrophic glycoproteins. The supernatant from these cells contain infectious virus, which can be used to infect the cells of interest.

Isolated recombinant polypeptides or proteins can be obtained by growing the described host cells under conditions that favor transcription and translation of the transfected nucleic acid. Recombinant

10 proteins produced by the transfected host cells are isolated using methods set forth herein and by methods well known to those skilled in the art.

Also provided are antibodies having specific reactivity with the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase protein or leukosialin protein of the present invention. Active fragments of antibodies, for example, Fab and Fab'$_2$ fragments, having specific reactivity with such proteins are intended to fall within the

15 definition of an "antibody." Antibodies exhibiting a titer of at least about 1.5×10^5, as determined by ELISA, are useful in the present invention.

The antibodies of the invention can be produced by any method known in the art. For example, polyclonal and monoclonal antibodies can be produced by methods described in Harlow and Lane, Antibodies: A Laboratory Manual (Cold Spring Harbor 1988), which is incorporated herein by reference. The

20 proteins, particularly core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase or leukosialin of the present invention can be used as immunogens to generate such antibodies. Altered antibodies, such as chimeric, humanized, CDR-grafted or bifunctional antibodies can also be produced by methods well known to those skilled in the art. Such antibodies can also be produced by hybridoma, chemical synthesis or recombinant methods described, for example, in Sambrook et al., supra, (1989).

25 The antibodies can be used for determining the presence or purification of the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase protein or the leukosialin protein of the present invention. With respect to the detecting of such proteins, the antibodies can be used for in vitro or in vivo methods well known to those skilled in the art.

Finally, kits useful for carrying out the methods of the invention are also provided. The kits can contain

30 a core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase protein, antibody or nucleic acid of the present invention and an ancillary reagent. Alternatively, the kit can contain a leukosialin protein, antibody or nucleic acid of the present invention and an ancillary reagent. An ancillary reagent may include diagnostic agents, signal detection systems, buffers, stabilizers, pharmaceutically acceptable carriers or other reagents and materials conventionally included in such kits.

35 A cDNA sequence encoding core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase was isolated and core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity was determined. This is the first report of transient expression cloning using CHO cells expressing polyoma large T antigen. The following examples are intended to illustrate but not limit the present invention.

40 **EXAMPLE I**

EXPRESSION CLONING IN COS-1 CELLS OF THE cDNA FOR THE PROTEIN CARRYING THE HEX-ASACCHARIDES

45 COS-1 cells were transfected with a cDNA library, pcDSRα-2F1, constructed from poly(A)$^+$ RNA of activated T lymphocytes, which express the core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase (Yokota et al., Proc. Natl. Acad. Sci. USA 83:5894-5898 (1986); Piller et al., supra, (1988), which are incorporated herein by reference). COS-1 cells support replication of the pcDSRα constructs, which contain the SV40 replication origin. Transfected cells were selected by panning using monoclonal antibody T305, which recognizes

50 sialylated branched hexasaccharides (Piller et al., supra, (1991); Saitoh et al., supra, (1991)). Methods referred to in this example are described in greater detail in the examples that follow.

Following several rounds of transfection, one plasmid, pcDSRα-leu, directing high expression of the T305 antigen was identified. The cloned cDNA insert was isolated and sequenced, then compared with other reported sequences. The newly isolated cDNA sequence was nearly identical to the sequence

55 reported for leukosialin, except the 5'-flanking sequences were different (Pallant et al., Proc. Natl. Acad. Sci. USA 86:1328-1332 (1989), which is incorporated herein by reference).

Comparison of the cloned cDNA sequence with the genomic leukosialin DNA sequence revealed the start site of the cDNA sequence is located 259 bp upstream of the transcription start site of the previously

reported sequence (Figure 2; compare Exon 1' and Exon 1) (Shelley et al., Biochem. J. 270:569-576 (1990); Kudo and Fukuda, J. Biol. Chem. 266:8483-8489 (1991), which are incorporated herein by reference). A consensus splice site was identified at the exon-intron junction of the newly identified 122 bp exon 1' in pcDSRα-leu (Breathnach and Chambon, Ann. Rev. Biochem. 50:349-383 (1981), which is incorporated
5 herein by reference). This splice site is followed by the exon 2 sequence.

These results indicate the T305 antibody preferentially binds to branched hexasaccharides attached to leukosialin. Indeed, a small amount of the hexasaccharides (approximately 8% of the total) was detected in O-glycans isolated from control COS-1 cells. T305 binding is similar to anti-M and anti-N antibodies, which recognize both the glycan and polypeptide portions of erythrocyte glycoprotein, glycophorin (Sadler et al.,
10 J. Biol. Chem 254: 2112-2119 (1979), which is incorporated herein by reference). These observations are consistent with reports that only leukosialin strongly reacted with T305 in Western blots of leukocyte cell extracts, even though leukocytes also express other glycoproteins, such as CD45, that must also contain the same hexasaccharides (Piller et al., supra, (1991); Saitoh et al., supra, (1991)).

15 **EXAMPLE II**

ESTABLISHMENT OF CHO CELL LINES THAT STABLY EXPRESS POLYOMA VIRUS LARGE T ANTI-GEN AND LEUKOSIALIN

20 T305 preferentially binds to branched hexasaccharides attached to leukosialin. Such hexasaccharides are not present on the erythropoietin glycoprotein produced in CHO cells, although the glycoprotein does contain the precursor tetrasaccharide (Sasaki et al., J. Biol. Chem. 262:12059-12076 (1987), which is incorporated herein by reference). T305 antigen also is not detectable in CHO cells transiently transfected with pcDSRα-leu. In order to screen for the presence of a cDNA clone expressing core 2 $\beta1{\rightarrow}6$ N-
25 acetylglucosaminyltransferase activity, a CHO cell line expressing both leukosialin and polyoma large T antigen was established (see, for example, Heffernan and Dennis Nucl. Acids Res. 19:85-92 (1991), which is incorporated herein by reference).

Vectors: A plasmid vector, pPSVE1-PyE, which contains the polyoma virus early genes under the control of the SV40 early promoter, was constructed using a modification of the method of Muller et al., Mol. Cell.
30 Biol. 4:2406-2412 (1984), which is incorporated herein by reference. Plasmid pPSVE1 was prepared using pPSG4 (American Type Culture Collection 37337) and SV40 viral DNA (Bethesda Research laboratories) essentially as described by Featherstone et al., Nucl. Acids Res. 12:7235-7249 (1984), which is incorporated herein by reference. Following EcoRI and HincII digestion of plasmid pPyLT-1 (American Type Culture Collection 41043), a DNA sequence containing the carboxy terminal coding region of polyoma virus
35 large T antigen was isolated. The HincII site was converted to an EcoRI site by blunt-end ligation of phosphorylated EcoRI linkers (Stratagene). Plasmid pPSVE1-PyE was generated by inserting the carboxy-terminal coding sequence for large T antigen into the unique EcoRI site of plasmid pPSVE1.

Plasmid pZIPNEO-leu was constructed by introducing the EcoRI fragment of PEER-3 cDNA, which contains the complete coding sequence for human leukosialin, into the unique EcoRI site of plasmid
40 pZIPNEO (Cepko et al., Cell 37:1053-1063 (1984), which is incorporated herein by reference). Plasmid structures were confirmed by restriction mapping and by sequencing the construction sites. pZIPNEO was kindly provided by Dr. Channing Der.

Transfection: CHODG44 cells were grown in 100 mm tissue culture plates. When the cells were 20% confluent, they were co-transfected with a 1:4 molar ratio of pZIPNEO-leu and pPSVE1-PyE using the
45 calcium phosphate technique (Graham and van der Eb, Virology 52:456-467 (1973), which is incorporated herein by reference). Transfected cells were isolated and maintained in medium containing 400 μg/ml G-418 (active drug).

Leukosialin expression: The total pool of G418-resistant transfectants was enriched for human leukosialin expressing cells by a one-step panning procedure using anti-leukosialin antibodies and goat anti-rabbit IgG
50 coated panning dishes (Sigma) (Carlsson and Fukuda J. Biol. Chem. 261:12779-12786 (1986), which is incorporated herein by reference). Clonal cell lines were obtained by limiting dilution. Six clonal cell lines expressing human leukosialin on the cell surface were identified by indirect immunofluorescence and isolated for further studies (Williams and Fukuda J. Cell Biol. 111:955-966 (1990), which is incorporated herein by reference).
55 **Polyoma virus-mediated replication:** The ability of the six clonal cell lines to support polyoma virus large T antigen-mediated replication of plasmids was assessed by determining the methylation status of transfected plasmids containing a polyoma virus origin of replication (Muller at al., supra, 1984; Heffernan and Dennis, supra, 1991). Plasmid pGT/hCG contains a fused $\beta1{\rightarrow}4$ galactosyltransferase and human

chorionic gonadotropin α-chain DNA sequence inserted in plasmid pcDNAI, which contains a polyoma virus replication origin (Aoki et al., Proc. Natl. Acad. Sci., USA 89, 4319-4323 (1992), which is incorporated herein by reference).

Plasmid pGT/hCG was isolated from methylase-positive E. coli strain MC1061/P3 (Invitrogen), which
5 methylates the adenine residues in the DpnI recognition site, "GATC". The methylated DpnI recognition site is susceptible to cleavage by DpnI. In contrast, the DpnI recognition site of plasmids replicated in mammalian cells is not methylated and, therefore, is resistant to DpnI digestion.

Methylated plasmid pGT/hCG was transfected by lipofection into each of the six selected clonal cell lines expressing leukosialin. After 64 hr, low molecular weight plasmid DNA was isolated from the cells
10 using the method of Hirt, J. Mol. Biol. 26:365-369 (1967), which is incorporated herein by reference. Isolated plasmid DNA was digested with XhoI and DpnI (Stratagene), subjected to electrophoresis in a 1% agarose gel, and transferred to nylon membranes (Micron Separations Inc., MA).

A 0.4 kb SmaI fragment of the β1→4 galactosyltransferase DNA sequence of pGT/hCG was radiolabel-ed with [^{32}P]dCTP using the random primer method (Feinberg and Vogelstein, Anal. Biochem. 132:6-13
15 (1983), which is incorporated herein by reference). Hybridization was performed using methods well-known to those skilled in the art (see, for example, Sambrook et al., supra, (1989)). Following hybridization, the membranes were washed several times, including a final high stringency wash in 0.1 x SSPE, 0.1% SDS for 1 hr at 65 °C, then exposed to Kodak X-AR film at -70 °C.

Four of the six clones tested supported replication of the pcDNAI-based plasmid, pGT/hCG (Fig. 3.A.,
20 lanes 1, 3, 4 and 5). MOP-8 cells, a 3T3 cell line transformed by polyoma virus early genes (Muller et al., supra, (1984)), expresses endogenous core 2 β1→6 N-acetylglucosaminyltransferase activity and was used as a control for the replication assay (Fig. 3.B., lane 1). One clonal cell line that supported pGT/hCG replication, CHO-Py-leu (Fig. 3.A., lane 5; Fig. 3.B., lanes 2 and 3) and expressed a significant amount of leukosialin, was selected for further studies. pGT/hCG was kindly provided by Dr. Michiko Fukuda.

25
EXAMPLE III

ISOLATION OF A cDNA SEQUENCE DIRECTING EXPRESSION OF THE HEXASACCHARIDE ON LEUKOSIALIN

30
Poly(A)$^+$ RNA was isolated from HL-60 promyelocytes, which contain a significant amount of the core 2 β1→6N-acetylglucosaminyltransferase (Saitoh et al., supra, (1991)). A cDNA expression library, pcDNAI-HL-60, was prepared (Invitrogen) and the library was screened for clones directing the expression of the T305 antigen.

35 Plasmid DNA from the pcDNAI-HL-60 cDNA library was transfected into CHO-Py-leu cells using a modification of the lipofection procedure, described below (Felgner et al., Proc. Natl. Acad. Sci. USA 84:7413-7417 (1987), which is incorporated herein by reference). CHO-Py-leu cells were grown in 100 mm tissue culture plates. When the cells were 20% confluent, they were washed twice with Opti-MEM I (GIBCO). Fifty μg of lipofectin reagent (Bethesda Research Laboratories) and 20 μg of purified plasmid
40 DNA were each diluted to 1.5 ml with Opti-MEM I, then mixed and added to the cells. After incubation for 6 hr at 37 °C, the medium was removed, 10 ml of complete medium was added and incubation was continued for 16 hr at 37 °C. The medium was then replaced with 10 ml of fresh medium.

Following a 64 hr period to allow transient expression of the transfected plasmids, the cells were detached in PBS/5mM EDTA, pH7.4, for 30 min at 37 °C, pooled, centrifuged and resuspended in cold
45 PBS/10mM EDTA/5% fetal calf serum, pH7.4, containing a 1:200 dilution of ascites fluid containing T305 monoclonal antibody. The cells were incubated on ice for 1 hr, then washed in the same buffer and panned on dishes coated with goat anti-mouse IgG (Sigma) (Wysocki and Sato Proc. Natl. Acad. Sci. USA 75:2844-2848 (1978); Seed & Aruffo Proc. Natl. Acad. Sci. USA 84:3365-3369 (1987), which are incorporated herein by reference). T305 monoclonal antibody was kindly provided by Dr. R.I. Fox, Scripps Research Founda-
50 tion, La Jolla, CA.

Plasmid DNA was recovered from adherent cells by the method of Hirt, supra, (1967), treated with DpnI to eliminate plasmids that had not replicated in transfected cells, and transformed into E. coli strain MC1061/P3. Plasmid DNA was then recovered and subjected to a second round of screening. E. coli transformants containing plasmids recovered from this second enrichment were plated to yield 8 pools of
55 approximately 500 colonies each. Replica plates were prepared using methods well-known to those skilled in the art (see, for example, Sambrook et al., supra, (1989)).

The pooled plasmid DNA was prepared from replica plates and transfected into CHO-Py-leu cells. The transfectants were screened by panning. One plasmid pool was selected and subjected to three subsequent

rounds of selection. One plasmid, pcDNAI-C2GnT, which directed the expression of the T305 antigen, was isolated. CHO-Py-leu cells transfected with pcDNAI-C2GnT express the antigen recognized by T305, whereas CHO-Py-leu cells transfected with pcDNAI are negative for T305 antigen (Fig. 4). These results show pcDNAI-C2GnT directs the expression of a new determinant on leukosialin that is recognized by T305 monoclonal antibody. This determinant is the branched hexasaccharide sequence, NeuNAcα2→3Galβ1→3-(NeuNAcα2→3Galβ1→4 GlcNAcβ1→6)GalNAc.

EXAMPLE IV

CHARACTERIZATION OF C2GnT

DNA sequence: The cDNA insert in plasmid pcDNAI-C2GnT was sequenced by the dideoxy chain termination method using Sequenase version 2 reagents (United States Biochemicals) (Sanger et al., Proc. Natl. Acad. Sci. USA 74:5463-5467 (1977), which is incorporated herein by reference). Both strands were sequenced using 17-mer synthetic oligonucleotides, which were synthesized as the sequence of the cDNA insert became known.

Plasmid pcDNAI-C2GnT contains a 2105 base pair insert (Fig. 5). The cDNA sequence (SEQ. ID. NO. 3) ends 1878 bp downstream of the putative translation start site. A polyadenylation signal is present at nucleotides 1694-1699. The significance of the large number of nucleotides between the polyadenylation signal and the beginning of the polyadenyl chain is not clear. However, this sequence is A/T rich.

Deduced amino acid sequence: The cDNA insert in plasmid pcDNAI-C2GnT encodes a single open reading frame in the sense orientation with respect to the pcDNAI promoter (Fig. 5). The open reading frame encodes a putative 428 amino acid protein having a molecular mass of 49,790 daltons.

Hydropathy analysis indicates the predicted protein is a type II transmembrane molecule, as are all previously reported mammalian glycosyltransferases (Schachter, supra, (1991)). In this topology, a nine amino acid cytoplasmic NH_2-terminal segment is followed by a 23 amino acid transmembrane domain flanked by basic amino acid residues. The large COOH-terminus consists of the stem and catalytic domains and presumably faces the lumen of the Golgi complex.

The putative protein contains three potential N-glycosylation sites (Fig. 5, asterisks). However, one of these sites contains a proline residue adjacent to asparagine and is not likely utilized in vivo.

No matches were obtained when the C2GnT cDNA sequence and deduced amino acid sequence were compared with sequences listed in the PC/Gene 6.6 data bank. In particular, no homology was revealed between the deduced amino acid sequence of C2GnT and other glycosyltransferases, including N-acetylglucosaminyltransferase I (Sarkar et al., Proc. Natl. Acad. Sci. USA 88:234-238 (1991), which is incorporated herein by reference).

mRNA expression: Poly(A)$^+$ RNA was prepared using a kit (Stratagene) and resolved by electrophoresis on a 1.2% agarose/2.2 M formaldehyde gel, and transferred to nylon membranes (Micro Separations Inc., MA) using methods well-known to those skilled in the art (see, for example, Sambrook et al., supra, (1989)). Membranes were probed using the EcoRI insert of pPROTA-C2GnT (see below) radiolabeled with [^{32}P]-dCTP by the random priming method (Feinberg and Vogelstein, supra, (1983). Hybridization was performed in buffers containing 50% formamide for 24 hr at 42°C (Sambrook et al., supra, (1989)). Following hybridization, filters were washed several times in 1xSSPE/0.1% SDS at room temperature and once in 0.1xSSPE/0.1% SDS at 42°C, then exposed to Kodak X-AR film at -70°C.

Fig. 6 compares the level of core 2 β1→6 N-acetylglucosaminyltransferase mRNA isolated from HL-60 promyelocytes, K562 erythroleukemia cells, and poorly metastatic SP and highly metastatic L4 colonic carcinoma cells. The major RNA species migrates at a size essentially identical to the ~2.1 kb C2GnT cDNA sequence. The same result is observed for HL-60 cells and the two colonic cell lines, which apparently synthesize the hexasaccharides. In addition, two transcripts of ~3.3 kb and 5.4 kb in size were detected in these cell lines. The two larger transcripts may result from differential usage of polyadenylation signals.

No hybridization occurred with poly(A)$^+$ RNA isolated from K562 cells, which lack the hexasaccharide, but synthesize the tetrasaccharide (Carlsson et al., supra, (1986)), which is incorporated herein by reference. Similarly, no hybridization was observed for poly(A)$^+$ RNA isolated from CHO-Py-leu cells (Fig. 6, lane 1).

EXAMPLE V

EXPRESSION OF ENZYMATICALLY ACTIVE $\beta 1 \rightarrow 6$ N-ACETYLGLUCOSAMINYLTRANSFERASE

5   In order to confirm that C2GnT cDNA encodes for core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase, enzymatic activity was examined in CHO-Py-leu cells transfected with pcDNAI or pcDNAI-C2GnT. Following a 64 hr period to allow transient expression, cell lysates were prepared and core 2 $\beta 1 \rightarrow 6$ N-acetyl-glucosaminyltransferase activity was measured.

N-acetylglucosaminyltransferase assays were performed essentially as described by Saitoh et al.,
10   supra, (1991), Yousefi et al., supra, (1991), and Lee et al., J. Biol. Chem. 265:20476-20487 (1990), which is incorporated herein by reference. Each reaction contained 50 mM MES, pH7.0, 0.5 μCi of UDP-[^{3}H]GlcNAc in 1 mM UDP-GlcNAc, 0.1 M GlcNAc, 10 mM Na$_2$EDTA, 1mM of acceptor and 25 μl of either cell lysate, cell supernatant or IgG-Sepharose matrix in a total reaction volume of 50 μl.

Reactions were incubated for 1 hr at 37°C, then processed by C18 Sep-Pak chromatography (Waters)
15   (Palcic et al., J. Biol. Chem. 265:6759-6769 (1990), which is incorporated herein by reference). Core 2 and core 4 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase were assayed using the acceptors p-nitrophenyl Gal$\beta 1 \rightarrow 3$Gal-NAc and p-nitrophenyl GlcNAc$\beta 1 \rightarrow 3$GalNAc, respectively (Toronto Research Chemicals).

UDP-GlcNAc:α-Man $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase(V) was assayed using the acceptor GlcNAc$\beta 1 \rightarrow 2$Man$\alpha 1 \rightarrow 6$Glc-β-O-(CH$_2$)$_7$CH$_3$.   The   blood   group   I   enzyme,   UDP-
20   GlcNAc:GlcNAc$\beta 1 \rightarrow 3$Gal$\beta 1 \rightarrow 4$GlcNAc (GlcNAc to Gal) $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase, was assayed using   GlcNAc$\beta 1 \rightarrow 3$Gal$\beta 1 \rightarrow 4$GlcNAc$\beta 1 \rightarrow 6$Man$\alpha 1 \rightarrow 6$Man$\beta 1 \rightarrow$O-(CH$_2$)$_8$COOCH$_3$ or Gal$\beta 1 \rightarrow 4$GlcNAc$\beta 1 \rightarrow 3$Gal$\beta 1 \rightarrow 4$GlcNAc$\beta 1 \rightarrow 3$Gal$\beta 1 \rightarrow 4$GlcNAc$\beta 1 \rightarrow$O-(CH$_2$)$_7$CH$_3$ as acceptors (Gu et al., J. Biol. Chem. 267:2994-2999 (1992), which is incorporated herein by reference). Synthetic acceptors were kindly provided by Dr. Olé Hindsgaul, University of Alberta, Canada.
25   Results of these assays are shown in Table I. Assuming transfection efficiency of the cells is approximately 20-30%, the level of enzymatic activity directed by cells transfected with pcDNAI-C2GnT is roughly equivalent to the level observed in HL-60 cells.

30

35

40

45

50

55

TABLE I

Core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase activity in CHO-Py-leu cell extracts transfected with pcDNAI or pcDNAI-C2GnT.	
Vector	Core 2 $\beta 1 \rightarrow 6$ GlcNAc transferase activity (pmol/mg of protein/hr)
pcDNAI	n.d.
pcDNAI-C2GnT	764

CHO-Py-leu cells were transfected with pcDNAI or pcDNAI-C2GnT, as described in the specification. Endogenous activity was measured in the absence of acceptor and subtracted from values determined in the presence of added acceptor. Gal$\beta 1 \rightarrow 3$GalNAcα-p-nitrophenyl was used as an acceptor. n.d. = not detectable. For comparison, the core 2 $\beta 1 \rightarrow 6$ N-acetylglucosaminyltransferase activity measured in HL-60 cells under identical conditions was 3228 pmol/mg of protein per hr.

In order to unequivocally establish that C2GnT cDNA sequence encodes core 2 $\beta1\rightarrow6$ N−acetyl-glucosaminyltransferase, plasmid, pPROTA-C2GnT was constructed containing the DNA sequence encoding the putative catalytic domain of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase fused in frame with the signal peptide and IgG binding domain of *S. aureus* protein A (Fig. 7). The putative catalytic domain is contained in a 1330 bp fragment of the C2GnT cDNA that encodes amino acid residues 38 to 428. Plasmid pPROTA was kindly provided by Dr. John B. Lowe.

The polymerase chain reaction (PCR) was used to insert EcoRI recognition sites on either side of the 1330 bp sequence in pcDNAI-C2GnT DNA. PCR was performed using the synthetic oligonucleotide primers 5'-TTTGAATTCCCCTGAATTTGTAAGTGTCAGACAC-3' (SEQ. ID. NO. 5) and 5'-TTTGAATTCGCAGAAACCATGCAGCTTCTCTGA-3' (SEQ. ID. NO. 6) (EcoRI recognition sites underlined). The EcoRI sites allowed direct, in-frame insertion of the fragment into the unique EcoRI site of plasmid pPROTA (Sanchez-Lopez et al., J. Biol. Chem. 263:11892-11899 (1988), which is incorporated herein by reference).

The nucleotide sequence of the insert as well as the proper orientation were confirmed by DNA sequencing using the primers described above for cDNA sequencing. Plasmid pPROTA-C2GnT allows secretion of the fusion protein from transfected cells and binding of the secreted fusion protein by insolubilized immunoglobulins.

Either pPROTA or pPROTA-C2GnT was transfected into COS-1 cells. Following a 64 hr period to allow transient expression, cell supernatants were collected (Kukowska-Latallo et al., supra, (1990)). Cell supernatants were cleared by centrifugation, adjusted to 0.05% Tween 20 and either assayed directly for core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity or used in IgG-Sepharose (Pharmacia) binding studies. For the latter assay, supernatants (10 ml) were incubated batchwise with approximately 300 μl of IgG-Sepharose for 4 hr at 4°C. The matrices were then extensively washed and used directly for glycosyltransferase assays.

No core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity was detected in the medium of COS-1 cells transfected with the control plasmid, pPROTA. Similarly, no enzymatic activity was associated with IgG-Sepharose beads. In contrast, a significant level of core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity was detected in the medium of COS-1 cells transfected with pPROTA-C2GnT. The activity also associated with the IgG-Sepharose beads (Table II). No activity was detected in the supernatant following incubation of the supernatant with IgG-Sepharose.

TABLE II

Determination of Enzymatic Activities Directed by pPROTA-C2GnT.

Acceptors and linkages formed	Radioactivity (cpm) with (+) and without (−) acceptor	
	−	+
GlcNAcß1 6 Galß1→3GalNAc (core 2-GnT)	109	1048
GlcNAcß1 6 GlcNAcß1→3GalNAc (core 4-GnT)	111	113
GlcNAcß1 6 GlcNAcß1→2Man (GnTV)	118	115
GlcNAcß1 6 GlcNAcß1→3Gal (I-GnT)	111	113
GlcNAcß1 6 Galß1→4GlcNAcß1→3Gal (I-GnT)	99	96

COS-1 cells were transfected with pPROTA-C2GnT and the conditioned media were incubated with IgG-Sepharose. The proteins bound to the IgG-Sepharose were assayed for ß1→6 N-acetylglucosaminyltransferase activity by using appropriate acceptors. The linkages formed are indicated by italics. Similar results were obtained in three independent experiments.

EXAMPLE VI

DETERMINATION OF C2GnT SPECIFICITY

5   Four types of $\beta1\rightarrow6$ N-acetylglucosaminyltransferase linkages have been reported, including core 2 and core 4 in O-glycans, I-antigen and a branch attached to mannose that forms tetraantennary N-glycans (see Table II). In order to determine whether these different structures are also synthesized by the cloned C2GnT cDNA sequence, enzymatic activity was determined using five different acceptors.

       As shown in Table II, the fusion protein was only active with the acceptor for core 2 formation. The
10   same was true when the formation of $\beta1\rightarrow6$ N-acetylglucosaminyl linkage to internal galactose residues was examined (Table II, see structure at bottom). This result precludes the likelihood that the enzyme encoded by the C2GnT cDNA sequence may add N-acetylglucosamine to a non-reducing terminal galactose. The HL-60 core 2 $\beta1\rightarrow6$ N-acetylglucosaminyltransferase is exclusively responsible for the formation of the GlcNAc$\beta1\rightarrow6$ branch on Gal$\beta1\rightarrow3$ GalNAc.

15   Although the invention has been described with reference to the disclosed embodiments, it should be understood that various modifications can be made without departing from the spirit of the invention. Accordingly, the invention is limited only by the following claims.

Lowe et al., Cell 63:475-484 (1990)

Brandley et al., Cell 63:861-863 (1990)

20   Phillips et al., Science 250:1130-1132 (1990)

Walz et al., Science 250:1132-1135 (1990)

Higgins et al., J. Biol. Chem. 266:6280-6290 (1991)

Schachter, Biochem. Cell Biol. 64:163-181 (1986)

25

30

35

40

45

50

55

SEQUENCE LISTING

(1) GENERAL INFORMATION:

    (i) APPLICANT:
        (A) NAME: La Jolla Cancer Research Foundation
        (B) STREET: 10901 North Torrey Pines Road
        (C) CITY: La Jolla
        (D) STATE: California
        (E) COUNTRY: U.S.A.
        (F) POSTAL CODE (ZIP): 92037

    (ii) TITLE OF INVENTION: A NOVEL BETA1-6
        N-ACETYLGLUCOSAMINYLTRANSFERASE, ITS ACCEPTOR MOLECULE,
        LEUKOSIALIN AND A METHOD FOR CLONING PROTEINS HAVING
        ENZYMATIC ACTIVITY

    (iii) NUMBER OF SEQUENCES: 8

    (iv) COMPUTER READABLE FORM:
        (A) MEDIUM TYPE: Floppy disk
        (B) COMPUTER: IBM PC compatible
        (C) OPERATING SYSTEM: PC-DOS/MS-DOS
        (D) SOFTWARE: PatentIn Release #1.0, Version #1.25 (EPO)

(2) INFORMATION FOR SEQ ID NO:1:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 900 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: both
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA (genomic)

    (ix) FEATURE:
        (A) NAME/KEY: CDS
        (B) LOCATION: 841..900

    (ix) FEATURE:
        (A) NAME/KEY: exon
        (B) LOCATION: 91..192
        (D) OTHER INFORMATION: /note= "EXON 1'IS LOCATED IN BOTH
            GENOMIC AND cDNA. IN THE cDNA EXON 1' IS
            IMMEDIATELY FOLLOWED BY EXON 2."

    (ix) FEATURE:
        (A) NAME/KEY: exon
        (B) LOCATION: 359..428
        (D) OTHER INFORMATION: /note= "EXON 1 IS LOCATED IN
            GENOMIC DNA"

    (ix) FEATURE:
        (A) NAME/KEY: intron
        (B) LOCATION: 193..806
        (D) OTHER INFORMATION: /note= "THIS SEGMENT OF NUCLEIC
            ACID CONSTITUTES INTRON SEQUENCE OF THE cDNA"

(ix) FEATURE:
    (A) NAME/KEY: exon
    (B) LOCATION: 807..900
    (D) OTHER INFORMATION: /note= "EXON 2 IS LOCATED IN BOTH
          GENOMIC AND cDNA. IN THE cDNA EXON 2 IMMEDIATELY
          FOLLOWS EXON 1'."

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
TTGGGGACCA CAAATGCAAA GGAAACCACC CTCCCCTCCC ACCTCCTCCT CTGCACCCTT      60

GAGTTCTCAG GCTCACATTC CCACCACCCA CCTCTGAGCC CAGCCCTCCC TAGCATCACC     120

ACTTCCATCC CATTCCTCAG CCAAGAGCCA GGAATCCTGA TTCCAGATCC CACGCTTCCC     180

TGCCTCCCTC AGGTGAGCCC CAGACCCCCA GGCACCCCGC TGGCCCCTGA AGGAGCAGGT     240

GATGGTGCTG TCTTCGCCCA GCAGCTGTGG GAGCAGGCGG GTGGGGCAGG ATGGAGGGGT     300

GGGTGGGGTG GGTGGAGCCA GGGCCCACTT CCTTTCCCCT TGGGGCCCTG TCCTTCCCAG     360

TCTTGCCCCA GCCTCGGGAG GTGGTGGAGT GACCTGGCCC CAGTGCTGCG TCCTTATCAG     420

CCGAGCCGGT AAGAGGGTGA GACTTGGTGG GGTAGGGGCC TCAGTGGGCC TGGGAATGTG     480

CCTGTGGCTT GAAAAGACTC TGACAGGTTA TGATGGGAAG AGATTGGGAG CCATTGGGCT     540

GCACAGGGTC AGGGAAGGCC AGGAGGGGCT GGTCACTGCT GGAATCTAAG CTGCTGAGGC     600

TGGAGGGAGC CTCAGGATGG GGCTGATGGG GGAGCTGCCA GCATCTGTTC CTCTGTCATT     660

TCTGATAACA GTAAAAGCCA GCATGGAAAA AACCGTTAAA CCGCAGGTTG GGCCTGGCCG     720

TTGGCAGGGA AGTGGGCAGA GGGGAGGCCC GGCCAGGTCC TCCGGCAACT CCCGCGTGTT     780

CTGCTTCTCC GGCTGCCCAC CTGCAGGTCC CAGCTCTTGC TCCTGCCTGT TTGCCTGGAA     840
```

```
ATG GCC ACG CTT CTC CTT CTC CTT GGG GTG CTG GTG GTA AGC CCA GAC      888
Met Ala Thr Leu Leu Leu Leu Leu Gly Val Leu Val Val Ser Pro Asp
 1               5                    10                   15
```

```
GCT CTG GGG AGC                                                      900
Ala Leu Gly Ser
            20
```

(2) INFORMATION FOR SEQ ID NO:2:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 20 amino acids
        (B) TYPE: amino acid
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: protein

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Met Ala Thr Leu Leu Leu Leu Leu Gly Val Leu Val Val Ser Pro Asp
 1               5                    10                   15

Ala Leu Gly Ser
            20
```

```
(2) INFORMATION FOR SEQ ID NO:3:

        (i) SEQUENCE CHARACTERISTICS:
              (A) LENGTH: 2105 base pairs
              (B) TYPE: nucleic acid
              (C) STRANDEDNESS: both
              (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: cDNA


       (ix) FEATURE:
              (A) NAME/KEY: CDS
              (B) LOCATION: 220..1504

       (ix) FEATURE:
              (A) NAME/KEY: polyA_signal
              (B) LOCATION: 1913..1918

       (ix) FEATURE:
              (A) NAME/KEY: misc_signal
              (B) LOCATION: 248..314
              (D) OTHER INFORMATION: /standard_name=
                      "SIGNAL/MEMBRANE-ANCHORING DOMAIN"


       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GTGAAGTGCT CAGAATGGGG CAGGATGTCA CCTGGAATCA GCACTAAGTG ATTCAGACTT        60

TCCTTACTTT TAAATGTGCT GCTCTTCATT TCAAGATGCC GTTGCAGCTC TGATAAATGC       120

AAACTGACAA CCTTCAAGGC CACGACGGAG GGAAAATCAT TGGTGCTTGG AGCATAGAAG       180

ACTGCCCTTC ACAAAGGAAA TCCCTGATTA TTGTTTGAA ATG CTG AGG ACG TTG         234
                                            Met Leu Arg Thr Leu
                                             1                   5

CTG CGA AGG AGA CTT TTT TCT TAT CCC ACC AAA TAC TAC TTT ATG GTT       282
Leu Arg Arg Arg Leu Phe Ser Tyr Pro Thr Lys Tyr Tyr Phe Met Val
                    10                  15                  20

CTT GTT TTA TCC CTA ATC ACC TTC TCC GTT TTA AGG ATT CAT CAA AAG       330
Leu Val Leu Ser Leu Ile Thr Phe Ser Val Leu Arg Ile His Gln Lys
                25                  30                  35

CCT GAA TTT GTA AGT GTC AGA CAC TTG GAG CTT GCT GGG GAG AAT CCT       378
Pro Glu Phe Val Ser Val Arg His Leu Glu Leu Ala Gly Glu Asn Pro
            40                  45                  50

AGT AGT GAT ATT AAT TGC ACC AAA GTT TTA CAG GGT GAT GTA AAT GAA       426
Ser Ser Asp Ile Asn Cys Thr Lys Val Leu Gln Gly Asp Val Asn Glu
        55                  60                  65

ATC CAA AAG GTA AAG CTT GAG ATC CTA ACA GTG AAA TTT AAA AAG CGC       474
Ile Gln Lys Val Lys Leu Glu Ile Leu Thr Val Lys Phe Lys Lys Arg
70                  75                  80                  85

CCT CGG TGG ACA CCT GAC GAC TAT ATA AAC ATG ACC AGT GAC TGT TCT       522
Pro Arg Trp Thr Pro Asp Asp Tyr Ile Asn Met Thr Ser Asp Cys Ser
                90                  95                  100
```

```
      TCT TTC ATC AAG AGA CGC AAA TAT ATT GTA GAA CCC CTT AGT AAA GAA            570
      Ser Phe Ile Lys Arg Arg Lys Tyr Ile Val Glu Pro Leu Ser Lys Glu
                  105             110                 115

5     GAG GCG GAG TTT CCA ATA GCA TAT TCT ATA GTG GTT CAT CAC AAG ATT            618
      Glu Ala Glu Phe Pro Ile Ala Tyr Ser Ile Val Val His His Lys Ile
                  120             125                 130

      GAA ATG CTT GAC AGG CTG CTG AGG GCC ATC TAT ATG CCT CAG AAT TTC            666
      Glu Met Leu Asp Arg Leu Leu Arg Ala Ile Tyr Met Pro Gln Asn Phe
10                135             140                 145

      TAT TGC GTT CAT GTG GAC ACA AAA TCC GAG GAT TCC TAT TTA GCT GCA            714
      Tyr Cys Val His Val Asp Thr Lys Ser Glu Asp Ser Tyr Leu Ala Ala
      150             155                 160                 165

15    GTG ATG GGC ATC GCT TCC TGT TTT AGT AAT GTC TTT GTG GCC AGC CGA            762
      Val Met Gly Ile Ala Ser Cys Phe Ser Asn Val Phe Val Ala Ser Arg
                      170             175                 180

      TTG GAG AGT GTG GTT TAT GCA TCG TGG AGC CGG GTT CAG GCT GAC CTC            810
      Leu Glu Ser Val Val Tyr Ala Ser Trp Ser Arg Val Gln Ala Asp Leu
                      185             190                 195
20
      AAC TGC ATG AAG GAT CTC TAT GCA ATG AGT GCA AAC TGG AAG TAC TTG            858
      Asn Cys Met Lys Asp Leu Tyr Ala Met Ser Ala Asn Trp Lys Tyr Leu
                      200             205                 210

      ATA AAT CTT TGT GGT ATG GAT TTT CCC ATT AAA ACC AAC CTA GAA ATT            906
      Ile Asn Leu Cys Gly Met Asp Phe Pro Ile Lys Thr Asn Leu Glu Ile
25                215             220                 225

      GTC AGG AAG CTC AAG TTG TTA ATG GGA GAA AAC AAC CTG GAA ACG GAG            954
      Val Arg Lys Leu Lys Leu Leu Met Gly Glu Asn Asn Leu Glu Thr Glu
      230             235                 240                 245

30    AGG ATG CCA TCC CAT AAA GAA GAA AGG TGG AAG AAG CGG TAT GAG GTC            1002
      Arg Met Pro Ser His Lys Glu Glu Arg Trp Lys Lys Arg Tyr Glu Val
                      250             255                 260

      GTT AAT GGA AAG CTG ACA AAC ACA GGG ACT GTC AAA ATG CTT CCT CCA            1050
      Val Asn Gly Lys Leu Thr Asn Thr Gly Thr Val Lys Met Leu Pro Pro
                      265             270                 275
35
      CTC GAA ACA CCT CTC TTT TCT GGC AGT GCC TAC TTC GTG GTC AGT AGG            1098
      Leu Glu Thr Pro Leu Phe Ser Gly Ser Ala Tyr Phe Val Val Ser Arg
                  280             285                 290

      GAG TAT GTG GGG TAT GTA CTA CAG AAT GAA AAA ATC CAA AAG TTG ATG            1146
40    Glu Tyr Val Gly Tyr Val Leu Gln Asn Glu Lys Ile Gln Lys Leu Met
                  295             300                 305

      GAG TGG GCA CAA GAC ACA TAC AGC CCT GAT GAG TAT CTC TGG GCC ACC            1194
      Glu Trp Ala Gln Asp Thr Tyr Ser Pro Asp Glu Tyr Leu Trp Ala Thr
      310             315                 320                 325

45    ATC CAA AGG ATT CCT GAA GTC CCG GGC TCA CTC CCT GCC AGC CAT AAG            1242
      Ile Gln Arg Ile Pro Glu Val Pro Gly Ser Leu Pro Ala Ser His Lys
                      330             335                 340

      TAT GAT CTA TCT GAC ATG CAA GCA GTT GCC AGG TTT GTC AAG TGG CAG            1290
      Tyr Asp Leu Ser Asp Met Gln Ala Val Ala Arg Phe Val Lys Trp Gln
50                345             350                 355
```

```
TAC TTT GAG GGT GAT GTT TCC AAG GGT GCT CCC TAC CCG CCC TGC GAT      1338
Tyr Phe Glu Gly Asp Val Ser Lys Gly Ala Pro Tyr Pro Pro Cys Asp
        360                 365                 370

GGA GTC CAT GTG CGC TCA GTG TGC ATT TTC GGA GCT GGT GAC TTG AAC      1386
Gly Val His Val Arg Ser Val Cys Ile Phe Gly Ala Gly Asp Leu Asn
        375                 380                 385

TGG ATG CTG CGC AAA CAC CAC TTG TTT GCC AAT AAG TTT GAC GTG GAT      1434
Trp Met Leu Arg Lys His His Leu Phe Ala Asn Lys Phe Asp Val Asp
390                 395                 400                 405

GTT GAC CTC TTT GCC ATC CAG TGT TTG GAT GAG CAT TTG AGA CAC AAA      1482
Val Asp Leu Phe Ala Ile Gln Cys Leu Asp Glu His Leu Arg His Lys
                410                 415                 420

GCT TTG GAG ACA TTA AAA CAC T GACCATTACG GGCAATTTTA TGAACAAGAA       1534
Ala Leu Glu Thr Leu Lys His
                425

GAAGGATACA CAAAACGTAC CTTATCTGTT TCCCCTTCCT TGTCAGCGTC GGGAAGATGG     1594

TATGAAGTCC TCTTTGGGGC AGGGACTCTA GTAGATCTTC TTGTCAGAGA AGCTGCATGG     1654

TTTCTGCAGA GCACAGTTAG CTAGAAAGGT GATAGCATTA AATGTTCATC TAGAGTTAAT     1714

AGTGGGAGGA GTAAAGGTAG CCTTGAGGCC AGAGCAGGTA GCAAGGCATT GTGGAAAGAG     1774

GGGACCAGGG TGGCTGGGGA AGAGGCCGAT GCATAAAGTC AGCCTGTTCC AAGTGCTCAG     1834

GGACTTAGCA AAATGAGAAG ATGTGACCTG TGCCAAAACT ATTTTGAGAA TTTTAAATGT     1894

GACCATTTTT CTGGTATGAA TAAACTTACA GCAACAAATA ATCAAAGATA CAATTAATCT     1954

GATATTATAT TTGTTGAAAT AGAAATTTGA TTGTACTATA AATGATTTTT GTAAATAATT     2014

TATATTCTGC TCTAATACTG TACTGTGTAG TGTGTCTCCG TATGTCATCT CAGGGAGCTT     2074

AAAATGGGCT TGATTTAACA TTGAAAAAAA A                                    2105
```

(2) INFORMATION FOR SEQ ID NO:4:

      (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 428 amino acids
          (B) TYPE: amino acid
          (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: protein

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
Met Leu Arg Thr Leu Leu Arg Arg Arg Leu Phe Ser Tyr Pro Thr Lys
 1               5                   10                  15

Tyr Tyr Phe Met Val Leu Val Leu Ser Leu Ile Thr Phe Ser Val Leu
            20                  25                  30

Arg Ile His Gln Lys Pro Glu Phe Val Ser Val Arg His Leu Glu Leu
            35                  40                  45

Ala Gly Glu Asn Pro Ser Ser Asp Ile Asn Cys Thr Lys Val Leu Gln
            50                  55                  60
```

```
        Gly Asp Val Asn Glu Ile Gln Lys Val Lys Leu Glu Ile Leu Thr Val
        65              70              75              80

5       Lys Phe Lys Lys Arg Pro Arg Trp Thr Pro Asp Asp Tyr Ile Asn Met
                        85              90              95

        Thr Ser Asp Cys Ser Ser Phe Ile Lys Arg Arg Lys Tyr Ile Val Glu
                    100             105             110

10      Pro Leu Ser Lys Glu Glu Ala Glu Phe Pro Ile Ala Tyr Ser Ile Val
                    115             120             125

        Val His His Lys Ile Glu Met Leu Asp Arg Leu Leu Arg Ala Ile Tyr
            130             135             140

15      Met Pro Gln Asn Phe Tyr Cys Val His Val Asp Thr Lys Ser Glu Asp
        145             150             155             160

        Ser Tyr Leu Ala Ala Val Met Gly Ile Ala Ser Cys Phe Ser Asn Val
                    165             170             175

20      Phe Val Ala Ser Arg Leu Glu Ser Val Val Tyr Ala Ser Trp Ser Arg
                    180             185             190

        Val Gln Ala Asp Leu Asn Cys Met Lys Asp Leu Tyr Ala Met Ser Ala
            195             200             205

25      Asn Trp Lys Tyr Leu Ile Asn Leu Cys Gly Met Asp Phe Pro Ile Lys
            210             215             220

        Thr Asn Leu Glu Ile Val Arg Lys Leu Lys Leu Leu Met Gly Glu Asn
        225             230             235             240

30      Asn Leu Glu Thr Glu Arg Met Pro Ser His Lys Glu Glu Arg Trp Lys
                    245             250             255

        Lys Arg Tyr Glu Val Val Asn Gly Lys Leu Thr Asn Thr Gly Thr Val
                    260             265             270

35      Lys Met Leu Pro Pro Leu Glu Thr Pro Leu Phe Ser Gly Ser Ala Tyr
                    275             280             285

        Phe Val Val Ser Arg Glu Tyr Val Gly Tyr Val Leu Gln Asn Glu Lys
            290             295             300

        Ile Gln Lys Leu Met Glu Trp Ala Gln Asp Thr Tyr Ser Pro Asp Glu
40      305             310             315             320

        Tyr Leu Trp Ala Thr Ile Gln Arg Ile Pro Glu Val Pro Gly Ser Leu
                    325             330             335

        Pro Ala Ser His Lys Tyr Asp Leu Ser Asp Met Gln Ala Val Ala Arg
                    340             345             350

45      Phe Val Lys Trp Gln Tyr Phe Glu Gly Asp Val Ser Lys Gly Ala Pro
                    355             360             365

        Tyr Pro Pro Cys Asp Gly Val His Val Arg Ser Val Cys Ile Phe Gly
            370             375             380

50      Ala Gly Asp Leu Asn Trp Met Leu Arg Lys His His Leu Phe Ala Asn
        385             390             395             400
```

```
Lys Phe Asp Val Asp Val Asp Leu Phe Ala Ile Gln Cys Leu Asp Glu
              405             410             415

His Leu Arg His Lys Ala Leu Glu Thr Leu Lys His
              420             425
```

(2) INFORMATION FOR SEQ ID NO:5:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 34 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: cDNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

TTTGAATTCC CCTGAATTTG TAAGTGTCAG ACAC   34

(2) INFORMATION FOR SEQ ID NO:6:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 33 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: cDNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

TTTGAATTCG CAGAAACCAT GCAGCTTCTC TGA   33

(2) INFORMATION FOR SEQ ID NO:7:

    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 15 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: protein

    (v) FRAGMENT TYPE: internal

    (ix) FEATURE:
        (A) NAME/KEY: CDS
        (B) LOCATION: 1..15
        (D) OTHER INFORMATION: /note= "PROTEIN A - C2GNT FUSION
             PROTEIN"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
GGG AAT TCC CCT GAA                                          15
Gly Asn Ser Pro Glu
  1           5
```

(2) INFORMATION FOR SEQ ID NO:8:

      (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 5 amino acids
          (B) TYPE: amino acid
          (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: protein

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
Gly Asn Ser Pro Glu
  1           5
```

Claims

1. A purified human protein or an active fragment thereof having $\beta1\rightarrow6$ N-acetylglucosaminyltransferase activity.

2. The purified protein of claim 1, wherein said activity is that of UDP-GlcNAc:Gal$\beta1\rightarrow3$GalNAc (GlcNAc to GalNAc) $\beta1\rightarrow6$ N-acetylglucosaminyltransferase.

3. The purified protein of claim 2, wherein said protein has a relative molecular weight of about 50 kD.

4. An isolated nucleic acid encoding the human protein or active fragment thereof of claim 1.

5. A vector containing the nucleic acid of claim 4.

6. The vector of claim 5, wherein said vector is a plasmid.

7. The vector of claim 5, wherein said vector is pcDNAI-C2GnT.

8. A host cell containing the vector of claim 5.

9. A purified human protein or a fragment thereof that is an acceptor molecule, said acceptor molecule being acted upon by the protein of claim 2 having activity which exclusively forms core 2 oligosaccharide structures in O-glycans.

10. The acceptor molecule of claim 9, wherein said acceptor molecule is leukosialin, CD43.

11. An isolated nucleic acid encoding the acceptor molecule of claim 9.

12. A vector containing the nucleic acid of claim 11.

13. The vector of claim 12, wherein said vector is a plasmid.

14. The vector of claim 12, wherein said vector is pcDSRα-leu.

15. A host cell containing the vector of claim 12.

16. A method of obtaining from a cell line, which does not normally contain a protein having catalytic activity or an acceptor molecule for said protein, a nucleic acid encoding said protein having catalytic activity comprising:

a. transfecting said cell line with a DNA sequence encoding the acceptor molecule, wherein the acceptor molecule is stably expressed in the cell line;

b. transfecting said cell line with a cDNA library containing said nucleic acid in a vector, wherein proteins encoded by the transfected cDNA are transiently expressed;

c. screening the transfected cells for expression of said protein having catalytic activity; and

d. isolating the nucleic acid encoding the protein having catalytic activity.

17. The vector of claim 16, wherein said vector replicates in the transfected cell line.

18. The vector in claim 17, wherein said vector is a plasmid.

19. The vector of claim 16, wherein said vector contains a viral replication origin.

20. The vector of claim 19, wherein said replication origin is the polyoma virus replication origin.
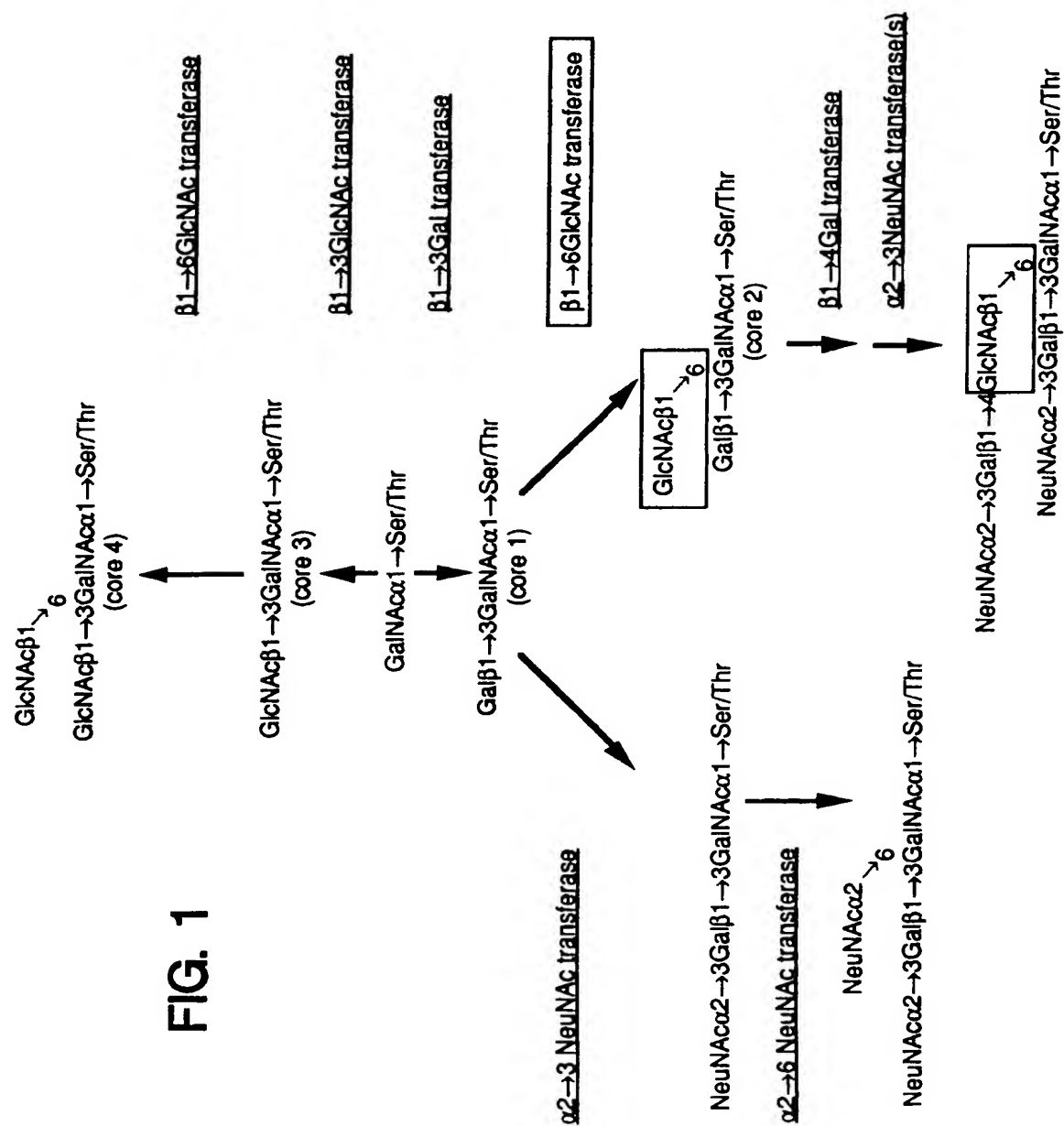
21. The cell line of claim 16, wherein said cell line supports replication of a vector.

22. The cell line of claim 16, wherein said cell line expresses polyoma virus large T antigen.

23. The cell line of claim 16, wherein said cell line is the Chinese hamster ovary cell line.

24. The cell line of claim 23, wherein said cell line is CHO-Py-leu.

25. A method of isolating a polypeptide having catalytic activity that forms core 2 oligosaccharide structures in *O*-glycans, said method comprising growing the host cell of claim 8 under conditions which favor expression of a nucleic acid encoding said polypeptide, and isolating said polypeptide so produced.

FIG. 1

```
        TTGGGGACCACAAATGCAAAGGAAACCACCCTCCCCTCCCACCTCCTCCTCTGCACCCTT              -299
        GAGTTCTCAGGCTCACATTCCACCACCACCTCTGAGCCCTCCCAGCATCACC                      -239
Exon1'  ACTTCCATCCCATTCCTCAGCCAAGAGCCAGGAATCCTGATTCCAGATCCCACGCTTCCC              -179
        TGCCTCCCTCAGGTGAGCCCCAGGCACCCGCTGGCCCTGAAGGAGAGCAGGT                      -119
        GATGGTGCTGTCTTCGCCCAGCAGCTGTGTGGGAGCAGGGGGTGGGGCAGGATGAGAGGGT            -59
        GGGTGGGGGTGGGAGCCAGGGCCCACTTCCTTTCCCCTTGGGGCCCTGTCCTTCCCGAG              2
Exon1   TCTTGCCCCAGCCTCGGGAGGAGGTGGTGGAGTGACCTGGCCCCAGTGCTGCGTCCTTATCAG          62
        CCGAGCCGGTAAGAGGGTGAGACTTGGTGGGGTAGGGGCCTCAGTGGGCCTGGGAATGTG            122
        CCTGTGGCTTGAAAAGACTCTGACAGGTTATGATGGGAAGAGATTGGGAGCCATTGGGCT            182
        GCACAGGGTCAGGGAAGGCCAGGAGGGGCTGGTCACTGCTGGAATCTAAGCTGCTGAGGC            242
        TGGAGGGAGCCTCAGGATGGGGCTGATGGGGGAGCTGCCAGCATCTGTTCCTCTGTCATT            302
        TCTGATAACAGTAAAAGCCAGCATGGAAAAAACCGTTAAACCGCAGGTTGGGCCTGGCCG            362
        TTGGCAGGGAAGTGGGCAGAGGGGAGGCCCGGCCAGGTCCTCCGGCAACTCCCGCGTGTT            422
        CTGCTTCTCCGGCTGCCCACCTGCAGGTCCCAGCTCTTGCCTCTGCTGTTTGCCTGGAA            482
Exon2   ATGGCCACGCTTCTCCTTCTCCTTGGGGTGCTGGTGGTAAGCCCAGACGCTCTGGGGAGC            542
        M  A  T  L  L  L  L  L  G  V  L  V  V  S  P  D  A  L  G  S              (20)
```
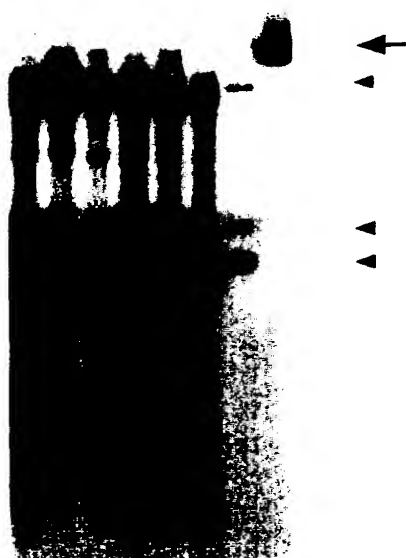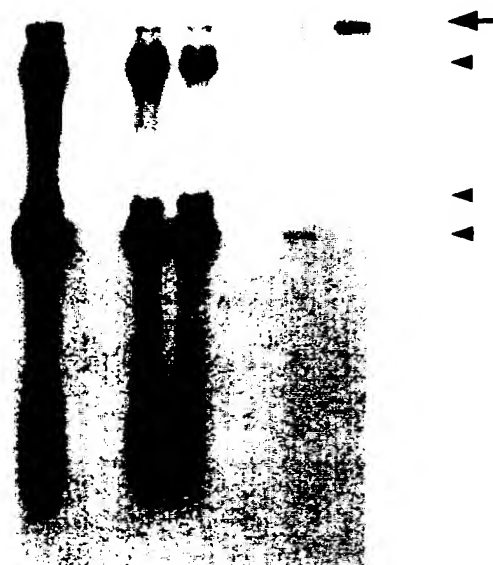
FIG. 2
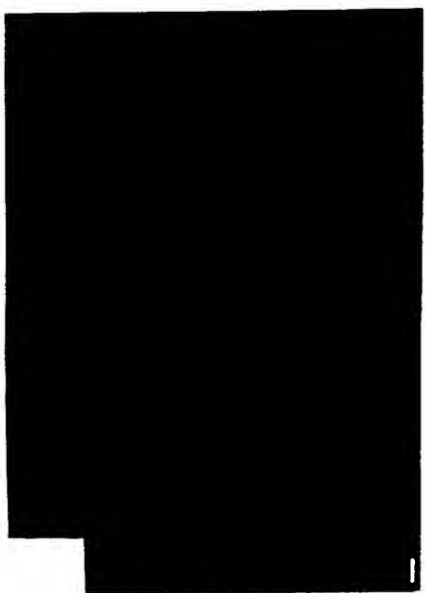
1 2 3 4 5 6 7 8

FIG. 3A

1   2 3   4 5

FIG. 3B
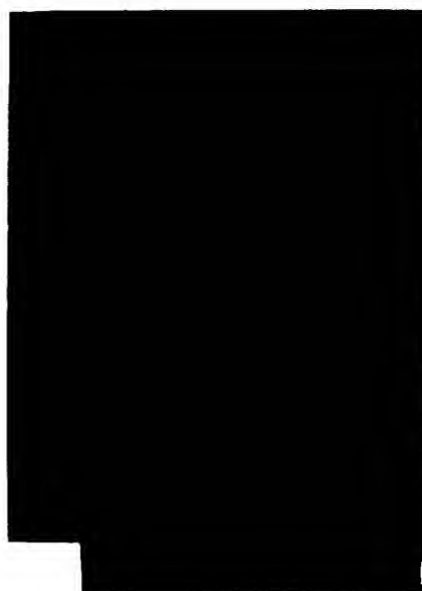
FIG. 4A

FIG. 4B

FIG. 4C

FIG. 4D

```
-219    GTGAAGTGCTCAGAATGGGCAGGATGTCACCTGAATCAGCACTAAGTGATTCAGACTTTCCTTACTTTTAAATGTGCTGCTCTTCATTTCAAGATGC                         -121

CGTTGCAGCTCTGATAAATGCAAACTGACAACCTTCAAGGCCACGACGGAGGAGGAAAATCATTGGTGCTTGGAGCATAGAGACTGCCCTTCACAAAGGAAATCCCTGATTATTGTTTGAA        -1

ATGCTGAGGACGTTGCTGCGAAGGAGACTTTTTCTTATCCACCAAATACTACTTTATGGTTCTTGTTTATCCCTAATCACCTTCTCCGTTTAAGGATTCATCAAAAGCCTGAATTT           120
M  L  R  T  L  L  R  R  L  F  S  Y  P  T  K  Y  Y  F  M  V  L  V  L  S  L  I  T  F  S  V  L  R  I  H  Q  K  P  E  F       40

GTAAGTGTCAGACACTTGGAGCTTGCTGGGGAGAATCCTAGTAGTGATATTAATTGCACCAAAGTTTACAGGGTGATGTAAATGAAATCCAAAAGGTAAAGCTTGAGATCCTAACAGTG       240
V  S  V  R  H  L  E  L  A  G  E  N  P  S  S  D  I  N  C  T  K  V  L  Q  G  D  V  N  E  I  Q  K  V  K  L  E  I  L  T  V       80
                              *

AAATTTAAAAAGCGCCCTCGGTGGACACCTGACGACTATATAAACATGACCAGTGACTGTTCTTCTTTCATCAAGAGACGCAAATATTGTAGAACCCTTAGTAAAGAAGAGGCGGAG       360
K  F  K  R  R  P  R  W  T  P  D  D  Y  I  N  M  T  S  D  C  S  S  F  I  K  R  R  K  Y  I  V  E  P  L  S  K  E  E  A  E      120
                                 *

TTTCCAATAGCATATTCTATAGTGGTTCATCACAAGATTGAAATGCTTGACAGGCTGCTGAGGGCCATCTATATGCCTCAGAATTTCATGTGCGTTCATGTGGACACAAAATCCGAGGAT     480
F  P  I  A  Y  S  I  V  V  H  H  K  I  E  M  L  D  R  L  L  R  A  I  Y  M  P  Q  N  F  Y  C  V  H  V  D  T  K  S  E  D      160

TCCTATTTAGCTGCAGTGATGGGCATCGCTTCCTGTTTTGTAATGTCTTTGGCCAGCCGATTGGAGAGTGTGGTTTATGCATGCTGGGATTGTGTGAGGGATCAAGAAGCCGAGACT       600
S  Y  L  A  A  V  M  G  I  A  S  C  F  S  N  V  F  V  A  S  R  L  E  S  V  V  Y  Y  A  S  W  S  R  V  Q  A  D  L  N  C  M    200

AAGGATCTCTATGCAATGAGTGCAAACTGGAAGTACTTGATAAATCTTTGTGGTATGGATTTCCCATTAAAACCAACCTAGAAATTGTCAGGAAGCTCAAGTTGTTAATGGGAGAAAAC     720
K  D  L  Y  A  M  S  A  N  W  K  Y  L  I  N  L  C  G  M  D  F  P  I  K  T  N  L  E  I  V  R  K  L  K  L  L  M  G  E  N      240
```

FIG. 5A

AACCTGGAAACGGAGGAGGATGCCATCCCATAAAGAGAAGAAAGGTGGAAGAAGCGGGTATGAGGTCGTTAATGGAAAATGCTTCCTCCACTCGAAACA 840
N L E T E R M P S H K E E R W K K R Y E V V N G K L T N T G T V K M L P P L E T 280

CCTCTCTTTTCTGGCAGTGCCTACTTCGTGGTCAGTAGGAGTATGTGGGTATTGTACTACAGAATGAAAAATCCAAAAGTTGATGGAGTGGCACAGACACATACAGCCCTGATGAG 960
P L F S G S A Y F V V S R E Y V G Y V L Q N E K I Q K L M E W A Q D T Y S P D E 320

TATCTCTGGGCCACCATCCAAAGGATTCCTGAAGTCCCGGGCTCACTCCCTGCCAGCCATAAGTAGTGATGATCTATCTGACATGCAAGCAGTTGCAGGTTTGTCAAGTGGCAGTACTTTGAG 1080
Y L W A T I Q R I P E V P G S L P A S H K Y D L S D M Q A V A R F V K W Q Y F E 360

GGTGATGTGTTCCAAGGGGTGCTCCCTACCCGGCCCTGCCGATGGAGTCCAGTGTCGCGTCTCAGTGTGCCATGTTTTCCGGAGCTGGTGACTTGAACTGGATGCTGCGCAAACACCACTTGTTTGCCAAT 1200
G D V S K G A P Y P P C D G V H V R S V C I F G A G D L N W M L R K H H L F A N 400

AAGTTTGACGTTGACGTGGATGTTGACCTCTTTGCCATCCAGTGTTGGATGGACATTTGAGACACAAAGCTTTGGAGACATTAAAACACTGACCATTACGGGCAATTTTATGAACAAGAGAAGG 1320
K F D V D V D L F A I Q C L D E H L R H K A L E T L K H end 428

ATACACACAAACGTACCTTATCTGTTCCCCTTCCTTGTCAGCGTCGGGAAGATGGTATGAAGTCCTCTTTGGGGCAGGACTCTAGTAGATCTTCTTGTCAGAGAAGCTGCATGGTTTCT 1440

GCAGAGCACAGTTAGCTAGAAAGGTGATAGCATTAAAATGTTCATCTCAGAGTTAATAGTGGGAGGAGTAAAGGTAGCCTTGAGGCCAGAGCAGTAGCAAGGCATTGTGAAAGAGGGGAC 1560

CAGGGTGGCTGGGGAAGAGAGGCCGATGCATAAAGTCAGCCTGTTCCAAGTGCTCAGGGGACTTAGCAAAAATGAGAAGATGACCTGTGCCAAAACTATTTTGAGAATTTAAATGTGACCA 1680

TTTTTCTGGTATG[AATAAA]CTTACAGCAACAAATAATCAAAGATACAATTAATCTGATATTATATTGTTGAAAATAGAAATTTGATTGTACTATAAATGATTTTTGTAAATAATTTATAT 1800

TCTGCTCTAATACTGTTACTGTGTAGTGTGTCTCCGTATGTCATCTCAGGGAGCTTAAAATGGGCTTGATTTAACATTGAAAAAAAA 1886
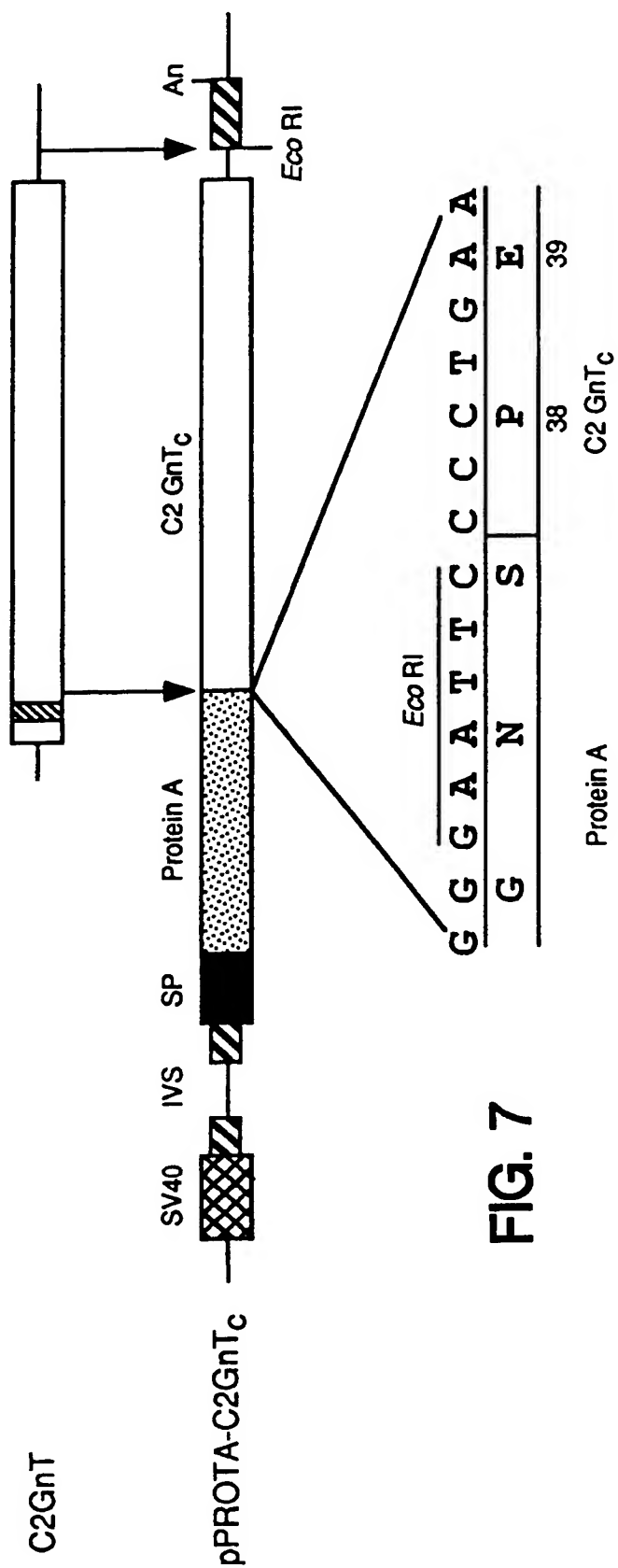
FIG. 5B

32

1   2   3   4   5

kb

7.46—

4.40—

2.37—

1.35—

0.24—

FIG. 6

FIG. 7